SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 1–18

THERE ARE ASYMMETRIC MINIMIZERS FOR THE ONE-DIMENSIONAL GINZBURG–LANDAU MODEL OF SUPERCONDUCTIVITY*

STUART P. HASTINGS† AND WILLIAM C. TROY†

Abstract. We study a boundary value problem associated with a system of two second-order differential equations with cubic nonlinearity which model a film of superconductor material subjected to a tangential magnetic field. We show that for an appropriate range of parameters there are *asymmetric* solutions and only trivial *symmetric* solutions. We then correct an error of the authors in [*Nonlinear Problems in Applied Mathematics*, SIAM, Philadelphia, PA, 1996, pp. 150–158] and show that the associated energy function is negative for the asymmetric solutions. Since the energy is zero for the trivial symmetric solution, it follows that a global minimizer of the energy is asymmetric. This property resolves a conjecture of Marcus [*Rev. Mod. Phys.*, 36 (1964), pp. 294–299].

Key words. superconductors, boundary value problem, minimizers, topological shooting

AMS subject classifications. 35Q35, 35B05

PII. S0036141096301956

1. Introduction.

In this paper we continue our recent studies [9], [10] of the one-dimensional Ginzburg–Landau model [8] for superconductors. Our main objective is to investigate the model for the existence of asymmetric minimizers of the appropriate energy integral. It is expected that the physically interesting solutions will be energy minimizers. These solutions satisfy a symmetric boundary value problem which was known to have a set of symmetric solutions. We will show, however, that in one space dimension, for some parameter values, the energy minimizer is an asymmetric solution to this symmetric boundary value problem.

The problem may be compared to studies in two dimensions, where the symmetry of the energy minimizer is unresolved. Recent papers on this include [12] and [14]. Work in two dimensions has largely dealt with a reduced problem in which variations in the magnetic field are ignored. In contrast, it is possible to include the magnetic field as a variable in the analysis of the one-dimensional model.

In order to properly describe our results we first need to give a brief development of the problem as well as a summary of our previous investigations.

In 1950 Ginzburg and Landau [8] proposed a model for the electromagnetic properties of a film of superconducting material of width d subjected to a tangential external magnetic field. Under the assumption that all quantities are functions only of the transverse coordinate, they proposed that the electromagnetic properties of the superconducting material are described by a pair $(\tilde{\phi}, \tilde{a})$, which minimizes the free energy functional

$$\mathcal{G} = \frac{1}{d}\int_{-d/2}^{d/2}\left(\tilde{\phi}^2(\tilde{\phi}^2-2) + \frac{2(\tilde{\phi}')^2}{k^2} + 2\tilde{\phi}^2\tilde{a}^2 + 2(\tilde{a}'-h_e)^2\right)dx. \tag{1.1}$$

The functional $\mathcal{G}$ is now known as the Ginzburg–Landau energy and provides a measure of the difference between normal and superconducting states of the material.

*Received by the editors December 17, 1996; accepted for publication (in revised form) January 2, 1998; published electronically September 26, 1998.
http://www.siam.org/journals/sima/30-1/30195.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (sph+@pitt.edu, troy@vms.cis.pitt.edu).

The variable $\tilde{\phi}$ is the order parameter which measures the density of superconducting electrons, and $\tilde{a}$ is the magnetic field potential. Also, h_e is the external magnetic field, and k is the dimensionless material constant distinguishing different superconductors, i.e., $0 < k < \frac{1}{\sqrt{2}}$ for type I superconductors and $k > \frac{1}{\sqrt{2}}$ for type II superconductors (see also [7]). The minimizer requirement, that $\mathcal{G}$ be stationary with respect to general first-order variations of the functions $\tilde{\phi}$ and $\tilde{a}$, leads to the boundary value problem

$$\tilde{\phi}'' = k^2\tilde{\phi}(\tilde{\phi}^2 + \tilde{a}^2 - 1), \tag{1.2}$$

$$\tilde{a}'' = \tilde{\phi}^2\tilde{a}, \tag{1.3}$$

$$\tilde{\phi}'\Big(\pm\frac{d}{2}\Big) = 0, \qquad \tilde{a}'\Big(\pm\frac{d}{2}\Big) = h_e. \tag{1.4}$$

It is routine to prove that $\mathcal{G}$ has a smooth minimizer satisfying (1.2)–(1.4) for any positive h_e. In 1964 Marcus [13] investigated the problem (1.2)–(1.4) and gave arguments which imply that a nontrivial minimizer of $\mathcal{G}$ should also satisfy

$$\tilde{\phi}(x) > 0 \quad \text{for all} \quad x \in \left[-\frac{d}{2}, \frac{d}{2}\right], \tag{1.5}$$

and, therefore, this is the only kind of solution we consider. A solution of (1.2)–(1.3) is called symmetric if

$$\tilde{\phi}'(0) = 0 \quad \text{and} \quad \tilde{a}(0) = 0. \tag{1.6}$$

It follows from (1.2) and (1.3) that if $\tilde{\phi}'(0) = \tilde{a}(0) = 0$, then $\tilde{\phi}$ is an even function and $\tilde{a}$ is odd. Thus $\tilde{\phi}$ is symmetric with respect to the origin and $\tilde{a}$ is antisymmetric. If (1.6) does not hold, then the solution is called asymmetric. Marcus makes the conjecture that a minimizer of $\mathcal{G}$ is probably a symmetric solution satisfying (1.2)–(1.6). However, he leaves open the possibility that asymmetric solutions may also exist.

In later work, Odeh [15] gave criteria for asymmetric solutions to exist by bifurcation as h_e increases, and in [2], [3], [4], [5], Bolley and Helffer give results implying that these criteria are satisfied for each $k > 0$. The existence of at least one symmetric solution has been investigated by Odeh [15], Wang and Yang [18], Yang [19], and also Bolley and Helffer [2], [3], [4], [5]. Numerical studies, such as the work of Seydel [16], [17] and also more recent theoretical work of Kwong [11], predict that a range of parameters exists for which asymmetric solutions and multiple symmetric solutions coexist. The work of Seydel [16] also predicts that there is a range of parameters for which no nontrivial symmetric solutions exist, yet asymmetric solutions do exist. In Figure 1 we show a solution found by Seydel. (The solution is rescaled from (1.2)–(1.4) by dividing $\tilde{\phi}$ through by $\tilde{\phi}(0)$.) Other papers, such as [6], considered the problem on an infinite interval, whereas in our work, the interval is large but finite. None of these studies addressed the physically important criterion of whether the solutions of the problem (1.2)–(1.5) are actually global minimizers of the energy functional $\mathcal{G}$. However, a recent paper by Aftalion [1] does discuss this problem and includes a conjecture that asymmetric solutions can have a lower energy than the symmetric solutions. This is confirmed in the current paper.

In two recent papers [9], [10] we began our investigation of the problem (1.2)–(1.5) with the goal of proving the existence of solutions predicted by the numerical studies

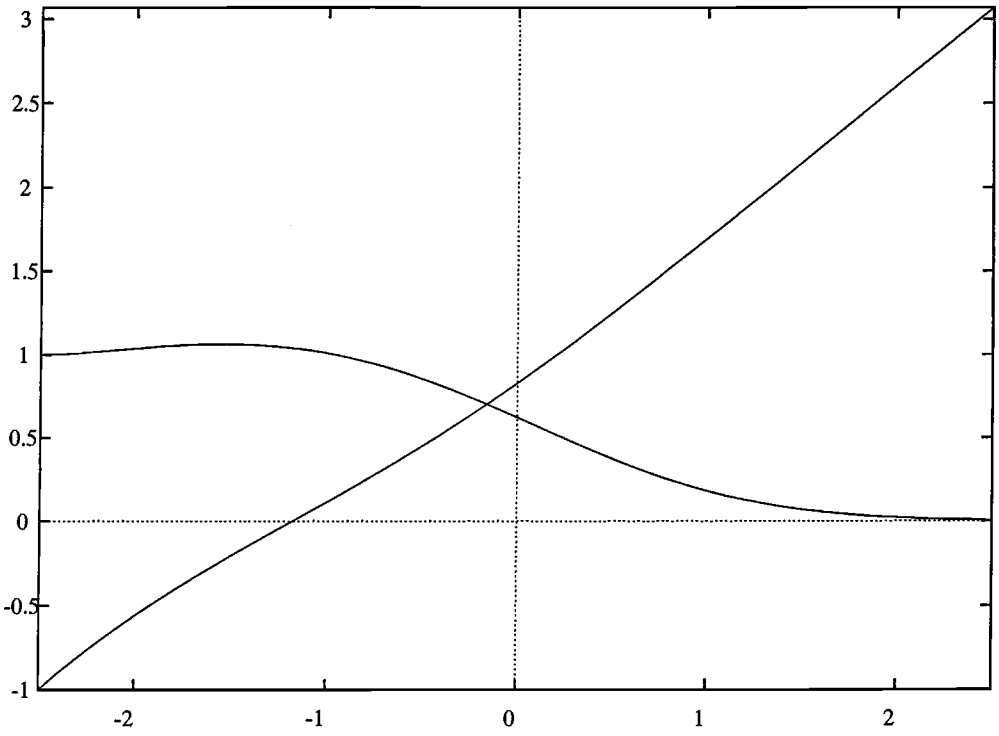


FIG. 1. *An asymmetric solution, showing* $\tilde{\phi}/\tilde{\phi}(0)$ *and* a*, found by Seydel. Here* $k = 1$.

described above. First, we studied the existence of multiple symmetric solutions and proved the following result.

THEOREM 1 (see [10]). (i) *Let* $k \in (0, \frac{1}{\sqrt{2}})$. *If* $d > 0$ *is sufficiently large, there is a range of values of* h_e *for which at least two symmetric solutions exist.*

(ii) *Let* $k > \frac{1}{\sqrt{2}}$. *If* $d > 0$ *is sufficiently large, there is a range of values of* h_e *for which at least three symmetric solutions exist.*

Remark. In related studies, Bolley and Helffer [2], [3], [4], [5] have investigated other properties of symmetric solutions, including bifurcation analysis and the uniqueness of solutions. Their analysis assumes that k tends to zero, whereas our results assume that $k > 0$ is fixed and d becomes large.

In [9] we shifted our attention from the symmetric solutions found in [10] to the study of asymmetric solutions. As mentioned above, the numerical experiments of Seydel (in particular, see Figure 6.10 in [16]) predict that there is a range of parameters in which there are no nontrivial symmetric solutions, yet asymmetric solutions do exist. This leaves open the possibility that in this parameter range the minimizer of $\mathcal{G}$ could be an asymmetric solution. Thus, our goal in [9] was to prove that there is a range of parameters in which only asymmetric solutions exist, and that the energy $\mathcal{G}$ is minimized by such solutions. The first step in proving this result is to find an upper bound on the values of h_e for which a symmetric solution exists. Thus, for fixed $k > 0$ and $d > 0$, we let h_e^* denote the supremum of the set of all positive h_e for which a *nontrivial* symmetric solution exists. Since we are assuming that $d \gg 1$ we define $h_e^{sym} = \overline{\lim}_{d\to\infty} h_e^*$.

In [9] we proved that if $k \geq \frac{1}{\sqrt{2.01}}$, then $h_e^{sym} < \sqrt{3}k$.

Unfortunately, a rescaling error led us to assert that this inequality was sufficient to prove that there are asymmetric minimizers. It turns out that we need a stronger estimate. We shall show below that the inequality

$$h_e^{sym} < 1.68k \tag{1.7}$$

suffices. This is obtained by a routine but tedious refinement of the proof in [9]. We shall not repeat that proof here. However, in the appendix we describe the changes which must be made in [9] to obtain (1.7). We add that the proof of (1.7) is considerably easier for large k. The proof for this case is in [9]. However, we want to include values in the range of type I superconductors, and this is more difficult.

Statement of main results. In this paper we will make use of (1.7) in proving that $\mathcal{G}$ has an asymmetric minimizer. We prove two main results. First, we fix $k \geq \frac{1}{\sqrt{2.01}}$ so that both type I and type II superconductors are included. There is no claim that this is a "best possible" estimate. It would be interesting to study the transition from asymmetric to symmetric minimizers as k decreases within the type I region. Then, in Theorem 2, we consider large d and prove that there exists a family of small amplitude asymmetric solutions of the problem (1.2)–(1.5). We will also prove that $\frac{h_e}{k} \geq 1.6831$ for large d, for each of the asymmetric solutions found in Theorem 2. This and (1.7) confirm the numerical prediction of Seydel in [16], that there is a parameter regime in which there are asymmetric solutions and only trivial symmetric solutions.

The work of Bolley and Helffer includes results which imply the existence of asymmetric solutions. The proofs, which must be pieced together from several papers, are by bifurcation theory, and do not appear to give the estimate of h_e which we obtain and which is essential in our discussion of whether these asymmetric solutions are energy minimizers.

The proof of Theorem 2 is given in section 2. Next, in Theorem 3 (proved in sections 3 and 4) we show that the energy is negative for the asymmetric solutions found in Theorem 2. This will allow us to prove that for large d a global minimizer of $\mathcal{G}$ must be asymmetric.

In addition to showing that asymmetric solutions exist and have small amplitude, we will prove that each of these has exactly one critical point (a relative maximum) in the open interval $(-\frac{d}{2}, \frac{d}{2})$ and that the relative maximum occurs at a value close to $-\frac{d}{2}$. Because of these properties, we find it convenient to rescale the problem. We introduce parameters r, h, m, and M and a new independent variable t by setting

$$r = \frac{1}{k^2},\ h = \frac{h_e}{k},\ m + M = kd,\ x = \frac{d}{2(m+M)}(2t + m - M). \tag{1.8}$$

Next, we define new dependent variables ψ and A by

$$\tilde{\phi}(x) = \beta\psi(t) \quad \text{and} \quad \tilde{a}(x) = A(t), \tag{1.9}$$

where $\beta = \tilde{\phi}(0)$. Then (1.1)–(1.5) become

$$\mathcal{G} = \frac{2\beta^2}{(m+M)} \int_{-m}^{M} \left(\psi^2 \left(\frac{\beta^2\psi^2}{2} - 1 + A^2 \right) + (\psi')^2 + \frac{1}{r\beta^2}(A' - h)^2 \right) dt, \tag{1.10}$$

$$\psi'' = \psi(\beta^2\psi^2 + A^2 - 1), \tag{1.11}$$

$$A'' = r\beta^2\psi^2 A, \tag{1.12}$$

$$\psi'(-m) = \psi'(M) = 0, A'(-m) = A'(M) = h > 0, \tag{1.13}$$

and

$$\psi > 0 \quad \text{in} \quad [-m, M]. \tag{1.14}$$

We now state our existence result in the following theorem.

THEOREM 2. *For sufficiently small $\beta > 0$ there is a value $h > 1.6831$ and a solution (ψ, A) of (1.11)–(1.14) defined on an interval $[-m, M]$ such that the following properties hold:*

(i) $m = m(\beta) > 0$, $M = M(\beta) > 0$, *and* $\lim_{\beta\to 0^+}(m, M) = (0, \infty)$;
(ii) $\psi' > 0$ *on* $(-m, 0)$, $\psi'(0) = 0$, *and* $\psi(0) = 1$;
(iii) $\psi' < 0$ *on* $(0, M)$;
(iv) *There is an* $h^0 > 1.6831$ *such that*

$$A'(-m) \to h^0 \quad \textit{as} \quad \beta \to 0^+. \tag{1.15}$$

Remark. Because of (1.8) and (1.9), each of the solutions (ψ, A) found in Theorem 2 corresponds to a solution $(\tilde\phi, \tilde a)$ of (1.2)–(1.5). Since ψ has a relative maximum at $t = 0$, (1.8) implies that $\tilde\phi'(x_{\max}) = 0$ where $x_{\max} = \frac{d}{2}\big(\frac{m-M}{m+M}\big)$. It follows from (1.9) and properties (i) and (iii) of Theorem 2 that for small $\beta > 0$, $x_{\max} < 0$, and $\tilde\phi' < 0$ for all $x \in (x_{\max}, \frac{d}{2})$. Therefore, $\tilde\phi'(0) < 0$, and (1.6) cannot hold. We conclude that $(\tilde\phi, \tilde a)$ is an asymmetric solution of (1.2)–(1.5).

We now state our second result. Recall that $\mathcal{G}$ gives the free energy of a solution.

THEOREM 3. *Let (ψ, A) be an asymmetric solution found in Theorem 2. There exists $\gamma > 0$ such that if $r \in (0, 2 + \gamma)$, then $\mathcal{G} < 0$ for sufficiently small $\beta > 0$.*

Asymmetric minimizers. We now return to the original system (1.1)–(1.5) and show that $\mathcal{G}$ has an asymmetric minimizer. Recall that $k \geq \frac{1}{\sqrt{2.01}}$ is fixed. Also, it follows from (1.7) that if $\frac{h_e}{k} \geq 1.6831$ and d is sufficiently large, then the only symmetric solution of (1.2)–(1.5) is the trivial solution $(\tilde\phi, \tilde a) = (0, h_e x)$, also known as the "normal state." Substitution of this pair into (1.1) shows that $\mathcal{G} = 0$. Next, we conclude from (1.8), (1.9), and the results of Theorems 2 and 3 that the problem (1.2)–(1.5) has an asymmetric solution for large d, that $\tilde a'(-\frac{d}{2}) \geq 1.6831\,k$, and that the corresponding energy $\mathcal{G}$ is negative. Therefore, in this parameter range, since $\mathcal{G}$ is zero for the trivial symmetric solution and negative for at least one asymmetric solution, a minimizer of $\mathcal{G}$ must be asymmetric.

2. Proof of Theorem 2. Our goal in this section is to show that for small $\beta > 0$ there is a solution (ψ, A) of the system

$$\psi'' = \psi(\beta^2\psi^2 + A^2 - 1), \tag{2.1a}$$

$$A'' = r\beta^2\psi^2 A \tag{2.1b}$$

on an interval $[-m, M]$, where $m > 0$ is small and $M > 0$ is large, and such that

$$\psi'(-m) = \psi'(M) = 0, \qquad A'(-m) = A'(M) = h > 0, \tag{2.2a}$$

$$\psi' > 0 \text{ on } (-m, 0), \qquad \psi' < 0 \text{ on } (0, M), \tag{2.2b}$$

$$\psi > 0 \text{ on } [-m, M], \tag{2.2c}$$

and

$$\psi(0) = 1, \qquad \psi'(0) = 0. \tag{2.2d}$$

Our method of proof uses a topological shooting argument. For this we begin by analyzing the important properties of solutions of the initial value problem (2.1a), (2.1b), (2.2d) when $\beta = 0$.

In this case A' is constant, and we set $A' = h$, where $h > 0$ is to be determined later. Setting $A(0) = A_0$, also to be determined later, we obtain the second-order linear equation

$$\psi'' = ((A_0 + ht)^2 - 1)\psi. \tag{2.3}$$

Because of (2.2d) we consider the solution of (2.3) such that the $\psi(0) = 1, \psi'(0) = 0$.

LEMMA 2.1. *Suppose that $-1 \le A_0 \le 0$. Then there is a unique $h_0 > 0$ (depending continuously on A_0) such that $\psi > 0, \psi' < 0$ on $(0, \infty)$, and $\psi(t) \to 0$ as $t \to \infty$. If $0 < h < h_0$, then $\psi = 0$ before $\psi' = 0$, while if $h > h_0$, then $\psi' = 0$ before $\psi = 0$.*

Proof. We consider the Riccati equation obtained by setting

$$\rho(s) = \frac{\psi'(s/h)}{\psi(s/h)}.$$

Then $\rho(0) = 0$ and

$$\rho' = \frac{H(s) - \rho^2}{h}, \tag{2.4}$$

where $H(s) = (A_0 + s)^2 - 1$. Since $A_0 \in [-1, 0], \rho$ initially decreases for any $h > 0$. Further, $\rho' < 0$ as long as $\rho(s)^2 > H(s)$. As long as $\rho' < 0$, the right side of (2.4) is an increasing and negative function of h and ρ. Suppose, for some first $s_0 > 0$, that $\rho'(s_0) = 0$. Then $H(s_0) = \rho(s_0)^2$ so that $A(s_0) > 1$. Therefore, $\rho''(s_0) = \frac{2A(s_0)}{h} > 0$. For $s > s_0$ it follows from the equation for ρ'' that $\rho'' > 0$ so that ρ increases until $\rho = 0$; i.e., $\psi' = 0$. Also, if ρ' becomes positive for some h_1, because ρ crosses the curve $\rho^2 = H(s)$, then the same must happen for any $h > h_1$. To see that there are values of h such that ρ' becomes positive we refer to (2.3). From that equation and our assumption that $\psi(0) = 1$, $\psi'(0) = 0$ we easily see that for large h, ψ' becomes positive before $\psi = 0$. This implies that ρ, and hence ρ', must become positive for large h. It then follows from continuity that the set of values of $h > 0$ such that this happens is open.

For small $h > 0$, on the other hand, we see that ρ decreases to below the curve $\rho = -s$. For example, if $-1 < A_0 < 0$ and $0 < h < 1 - A_0^2$, then $\rho'(0) < -1$, so immediately ρ decreases below $-s$. If $A_0 = -1$, then $\rho(0) = \rho'(0) = 0$. However, $\rho''(s) < \frac{2s-2}{h}$ and this integrates to show that for small h, $\rho(2) < -2$ and $\rho' < 0$ over $(0, 2]$. It follows from (2.4), and our assumption that $A_0 \in [-1, 0]$, that ρ' continues to decrease until $\rho \to -\infty$ at some finite s_1, that is, $\psi(s_1) = 0$. It is clear that the set of $h's$ such that $\rho(s) < -s$ for some s is an open subset of the interval $0 < h < \infty$.

Similarly, as we pointed out above, the set of $h's$ such that $\rho(s)^2 < H(s)$ for some s is also open.

Moreover, if ρ ever falls below $-s$, then we see from the equation for ρ'' that thereafter, $\rho' < -1$, and ρ decreases monotonically to $-\infty$. On the other hand, if ρ' is ever positive, then we see that ρ becomes positive, and because of (2.4) it must remain positive.

Hence, there is at least one h_0 such that on $0 \le s < \infty$, $\rho' \le 0$, and $0 > \rho > -s$. These bounds imply that the solution exists on the entire interval $[0, \infty)$. In order to complete the proof of Lemma 2.1 we need further properties of this solution. These are given in the following result.

LEMMA 2.2. *For such an* $h_0, \rho' < 0$, $-s < \rho(s)$ *on* $(0, \infty), \rho(s) < -\sqrt{H(s)}$ *if* $H(s) \ge 0$, *and* $\rho + \sqrt{H(s)} \to 0$ *as* $s \to \infty$.

Proof. The inequalities have already been proved. To see the limiting behavior of $\rho(s)$, we first note that for $h = h_0, \rho'(s)$ is bounded. If ρ' is unbounded, then it must get arbitrarily large and negative, but then the equation for ρ'' shows that ρ' remains large and negative and $\rho(s) < -s$ for some s.

Writing $\rho' = \frac{(\sqrt{H}-\rho)(\sqrt{H}+\rho)}{h}$ and noting that the first of these factors is unbounded, we see that $\sqrt{H(s)} + \rho(s)$ tends to zero, which implies Lemma 2.2.

To prove uniqueness we suppose that there is a second positive value of $h, h_1 < h_0$ for which this behavior occurs. Let ρ_1, ρ_0 denote the corresponding solutions. Then $\rho_1(0) = \rho_0(0) = 0$, and it is easily shown that

$$\frac{d}{ds}(\rho_1 - \rho_0) < 0$$

for all $s > 0$. Thus $\rho_1 - \rho_0$ could not approach zero as $s \to \infty$. However, Lemma 2.2 implies that $\rho_1 - \rho_0 = (\rho_1 + \sqrt{H}) - (\rho_0 + \sqrt{H}) \to 0$ as $s \to \infty$, a contradiction. Therefore, h_0 is unique.

To complete the proof of Lemma 2.1 we must show that h_0 depends continuously on A_0. This follows from the uniqueness of h by a standard argument.

Now let $A_0 = -1$, and define h^0 from Lemma 2.1 with $A_0 = -1$. We will find the desired solutions near the point $(A_0, h) = (-1, h^0)$.

LEMMA 2.3. $h^0 > 1.6831$.

Proof. Let

$$\hat{\rho}(t) = \frac{\psi'(t)}{\psi(t)},$$

so that

$$\hat{\rho}' = H(ht) - \hat{\rho}^2. \tag{2.5}$$

Here, because $A_0 = -1$, $H(s) = -2s + s^2$, where $s = ht$ as before. We use the following result about a solution $\hat{\rho}$ of (2.5) such that $\hat{\rho}(0) = 0$. If $\hat{\rho}''' < 0$ for some t_1, with $\hat{\rho}, \hat{\rho}', \hat{\rho}'' < 0$ on $(0, t_1)$, then $\hat{\rho}$ decreases monotonically to $-\infty$. To see this, compute the equation satisfied by $\hat{\rho}''''$, that is,

$$\hat{\rho}'''' + 2\hat{\rho}\hat{\rho}''' = -6\hat{\rho}'\hat{\rho}''.$$

If there were a first $y > t_1$ where $\hat{\rho}'''(y) = 0$, then $\hat{\rho}''''(y) \ge 0$. However, the last equation gives $\hat{\rho}''''(y) < 0$ since the definition of y implies that $\hat{\rho}'\hat{\rho}''$ is positive and

increasing on (t_1, y). From these derivative properties it follows that $\hat{\rho}$ decreases below $-s$, and, hence, $\hat{\rho}(s) \to -\infty$ at a finite value of s. Next, we define a sequence $\{\hat{\rho}_N\}$ of functions, all defined on $[0, 2]$ as follows:

$$\hat{\rho}_0(t) = \int_0^t (h^2 r^2 - 2hr)dr, \quad \hat{\rho}_{N+1}(t) = \hat{\rho}_0(t) - \int_0^t \hat{\rho}_N(r)^2 dr \quad \text{for} \quad N \geq 1.$$

From this definition it is evident that $\{\hat{\rho}_N\}$ forms a decreasing sequence of functions defined on [0,2], for all $N \geq 1$. Furthermore, it is easily shown that our solution $\hat{\rho}(t)$ satisfies $\hat{\rho}^{'''} < \hat{\rho}_i^{'''}$ on [0,2], for all $i \geq 1$. Setting $h = \frac{16831}{1000}$, we used computer algebra program Maple to compute $\hat{\rho}_6$, a polynomial of degree 225. All calculations are with integers and rational numbers so that there are no roundoff errors. From $\hat{\rho}_6$ and its first two derivatives we can compute $\hat{\rho}_7'''$ without having to compute $\hat{\rho}_7$. We find that $\hat{\rho}_7'''(6/5) < 0$ and that on [0,6/5], $\hat{\rho}, \hat{\rho}', \hat{\rho}^{''}$ are all negative. This completes the proof of the lemma.

Our solution (ψ, A) to (2.1a)–(2.2d) is obtained by perturbing $A(0)$ from -1, keeping $\beta = 0$, and then letting β be positive. Our argument is by "shooting," rather than by the use of bifurcation theory.

Thus, we will assume that

$$A(0) = -1 + \epsilon$$

for small $\epsilon \geq 0$, and we let $h_0 = h_0(\epsilon)$ denote the value of h found in Lemma 2.1.

In constructing solutions to (2.1)–(2.2), in order to get something meaningful at $\beta = 0$, we replace the boundary conditions on A in (2.2a) with

$$\int_{-m}^{M} \psi^2(t) A(t) dt = 0.$$

LEMMA 2.4. *Suppose that $\epsilon = 0$, (i.e., $A_0 = -1$) and let ψ_0 be the solution found in Lemma 2.1 where $h = h^0 = h_0(0)$. Then*

$$\int_0^\infty \psi_0(t)^2(-1 + h^0 t)dt = 0.$$

Proof. It is easily seen that $\psi_0(t) \to 0$ exponentially fast as $t \to \infty$. Thus, the integral in the lemma converges and $\lim_{t\to\infty}(-1 + h^0 t)^2 \psi_0(t)^2 = 0$. We set $\psi = \psi_0$ in (2.3), multiply by ψ_0', and integrate by parts to obtain the result.

From Lemma 2.1 and the definition of $h_0(\epsilon)$ we see that for each $h > h_0(\epsilon)$, there is a first $t = t_h > 0$ such that $\psi'(t_h) = 0, \psi' < 0$ and $\psi > 0$ on $(0, t_h)$, with $\psi(t_h) > 0$. By the implicit function theorem, t_h is continuous in h on $(h_0(\epsilon), \infty)$ since $\psi^{''}(t_h) > 0$. (If $\psi^{''}(t_h) = 0$, then $t_h = \frac{2}{h}$ and $\psi^{'''}(t_h) > 0$, a contradiction.) Further, $t_h \to \infty$ as $h \to h_0(\epsilon)$ from above.

Now consider small $\epsilon > 0$, set $h = h_0(\epsilon)$, and compute $\psi^{''}(0)$ and $\psi^{'''}(0)$. We find that $\psi''(0) = -2\epsilon + \epsilon^2$ and $\psi^{'''}(0) = 2(-1 + \epsilon)h_0(\epsilon)$. From this and the fact that $h_0(\epsilon) \to h^0 > 0$ as $\epsilon \to 0^+$, we see that for small $\epsilon, \psi'(-m) = 0$ for some $m = m_0(\epsilon) > 0$. Furthermore,

$$-m = \frac{-2\epsilon}{h_0(\epsilon)} + 0(\epsilon^2). \tag{2.6}$$

Here, $O(\epsilon^2) < L\epsilon^2$ for some L independent of ϵ. We now compute

$$I(\epsilon) = \int_{-m}^{\infty} \psi^2 A ds,$$

where $A(s) = -1+\epsilon+h_0(\epsilon)s$. Again, multiplying (2.3) by ψ' and integrating by parts give

$$0 = -\psi(-m)^2 A(-m)^2 + \psi(-m)^2 - 2h_0 \int_{-m}^{\infty} \psi(s)^2 A(s) ds.$$

Using (2.6) and the fact that $\psi(-m) = 1+O(\epsilon)$, we find that $I(\epsilon) = -\frac{\epsilon}{h_0} + O(\epsilon^2)$ as $\epsilon \to 0^+$. Thus, we have shown that with $h = h_0(\epsilon)$ for small ϵ, $I(\epsilon) < 0$. Fix $\epsilon > 0$ small enough so that this inequality holds for $h = h_0(\epsilon)$. Then for $h - h_0(\epsilon)$ positive but sufficiently small, $-m = -m(\epsilon, h)$ will still be defined as the largest negative zero of ψ', A is positive on (t_h, ∞), and therefore,

$$\int_{-m}^{t_h} \psi^2 A ds < 0.$$

Our goal now is to find an h and ϵ such that $-m$ and t_h are defined as the negative and positive zeros of ψ' closest to $t = 0$, with m small and t_h large, and such that

$$I = \int_{-m}^{t_h} \psi^2 A ds > 0.$$

We will see that it is not necessary to have the same m and t_h as before. Starting with $\epsilon = 0$ and $h = h_0(\epsilon)$, we now raise h, instead of ϵ. Thus, for small $h - h^0 > 0$ our solution satisfies $A(0) = -1$, $A'(0) = h > h^0 = h_0(0)$, $\psi'(t_h) = 0$ for some large t_h. In this case, we multiply (2.3) by ψ' and integrate from 0 to t_h, where t_h is the first positive zero of ψ'. Using integration by parts once again, we find that $I > 0$. We then keep h fixed and raise ϵ slightly, whereupon both $m = m(\epsilon, h)$ and t_h are defined, still with $I > 0$. We summarize these results in the following lemma.

LEMMA 2.5. *For each sufficiently small $\epsilon > 0$ there is an $h_1(\epsilon) > h_0(\epsilon)$ such that if $h_0(\epsilon) < h < h_1(\epsilon)$, then the solution of* (2.3) *with $A_0 = -1+\epsilon, \psi(0) = 1$, and $\psi'(0) = 0$ is decreasing on an interval $[0, M]$, increasing on an interval $[-m, 0]$, and $\psi'(-m) = \psi'(M) = 0$, and*

$$I(\epsilon, h) = \int_{-m}^{M} \psi^2 A ds < 0.$$

Furthermore, $\psi > 0$ on $[-m, M]$ and $h_1(\epsilon) \to h_0(0) = h^0$ as $\epsilon \to 0^+$. On the other hand, for each $h > h_0(0)$ sufficiently close to $h_0(0)$, there is an interval $(0, \epsilon_1(h))$ of $\epsilon's$ such that ψ has the same behavior, but $I(\epsilon, h) > 0$. As ϵ and $h - h_0(0)$ tend to zero, $m \to 0$ and $M \to \infty$.

COROLLARY 2.6. *For $\epsilon > 0$ sufficiently small, there is an $h_2 = h_2(\epsilon)$ such that if ψ is the solution of* (2.3) *with $A_0 = -1+\epsilon$ and $(\psi(0), \psi'(0)) = (1, 0)$, then there are values $M > m > 0$ such that*

$$\psi'(-m) = \psi'(M) = 0, \quad \psi' > 0 \text{ on } (-m, 0), \quad \psi' < 0 \text{ on } (0, M), \tag{2.7a}$$

$$\psi''(-m) \neq 0, \psi''(M) \neq 0. \tag{2.7b}$$

Further, $I(\epsilon, h_1(\epsilon)) < 0$ and $I(\epsilon, h_2(\epsilon)) > 0$. Finally, $h_2(\epsilon) \to h_0(0) = h^0$ as $\epsilon \to 0$.

With $\epsilon > 0$ sufficiently small so that Lemma 2.5 and Corollary 2.6 hold, we now raise β and consider solutions of (2.1) satisfying (2.2d). Since $\psi'' \neq 0$ at the zeros of ψ' when $\beta = 0$, it follows from the implicit function theorem that the zeros $-m < 0$ and $M > 0$ of ψ' persist as continuous functions of β and the values of $A(0)$ and $A'(0)$. For each sufficiently small $\epsilon > 0$ there exist functions $h_1(\epsilon)$ and $h_2(\epsilon)$ independent of β, such that for sufficiently small β (depending on ϵ) the solution (ψ, A) of (2.1) and (2.2d) with

$$A(0) = -1 + \epsilon, \quad A'(0) = h_1(\epsilon), \tag{2.8}$$

satisfies (2.7) on an interval $[-m, M]$, and

$$\int_{-m}^{M} \psi^2 A dt < 0.$$

Also, the solution of (2.1), (2.2d) with $A(0) = -1 + \epsilon$, $A'(0) = h_2(\epsilon)$ satisfies

$$\int_{-m}^{M} \psi^2 A dt > 0.$$

Furthermore, $\beta \to 0$ as $\epsilon \to 0$, and

$$(m, M) \to (0, \infty), \quad (h_1(\epsilon), h_2(\epsilon)) \to (h^0, h^0).$$

In addition, it follows from (2.1b) that A'' is uniformly bounded on $[-m, 0]$, and therefore, $A'(-m) = h \to h^0$ as $\epsilon \to 0$ and $\beta \to 0^+$. It then follows from continuity that for given fixed small ϵ, and sufficiently small $\beta \geq 0$, there is an $h \in (h_1(\epsilon), h_2(\epsilon))$ such that the solution $(\psi_{\epsilon,\beta}, A_{\epsilon,\beta})$ of (2.1), (2.2d), and (2.8) with $A'(-m) = h$ also satisfies $\int_{-m}^{M} \psi^2 A dt = 0$, so $A'(M) = h$. The conclusion of Theorem 2 now follows from the transformation (1.8)–(1.9).

In the next section we will turn to the proof of Theorem 3. For this we will use the following result, which is based on the construction given above. Let

$$\mathcal{E}(\psi, A) = \int_{-m}^{M} F(\psi, A)(t) dt,$$

where

$$F(\psi, A)(t) = r\left(\int_t^M \psi(s)^2 A(s) ds \right)^2 - \frac{\psi(t)^4}{2}.$$

LEMMA 2.7. *With $(\psi, A) = (\psi_{\epsilon,\beta}, A_{\epsilon,\beta})$ chosen as in the proof of Theorem 2, we have*

$$\lim_{\epsilon \to 0} \{ \lim_{\beta \to 0} \mathcal{E}(\psi, A) \} = \mathcal{E}(\psi_0, A^0),$$

where $A^0(s) = -1 + h^0 s$.

Proof. Our proof of Theorem 2 shows that

$$\lim_{\epsilon \to 0} \{ \lim_{\beta \to 0} (\psi_{\epsilon,\beta}(s), A_{\epsilon,\beta}(s)) \} = (\psi_0(s), A^0(s))$$

uniformly on compact intervals. Let $\delta > 0$ be given. We can choose $K_1 > 0$ such that

$$r\psi(K_1)^4 < \delta, \int_{K_1}^{\infty} \big|F(\psi_0, A^0)(t)\big|dt < \delta, \tag{2.9}$$

and

$$A^0(s)^2 - 2 > \frac{A^0(s)^2}{2} \tag{2.10}$$

for $s \geq K_1$. Then for sufficiently small $\epsilon > 0$ we choose $\beta_1 = \beta_1(\epsilon) > 0$ such that for $0 < \beta < \beta_1$,

$$\Big| \int_0^{K_1} (F(\psi, A)(t) - F(\psi_0, A^0)(t))dt \Big| < \delta. \tag{2.11}$$

Further, we can ensure that for $0 < \beta < \beta_1$, (2.10) also holds on $[K_1, M]$ with A substituted for A_0.

We consider two cases:

$$\frac{\psi_0'(K_1)}{\psi_0(K_1)} > -1 \tag{i}$$

and

$$\frac{\psi_0'(K_1)}{\psi_0(K_1)} \leq -1. \tag{ii}$$

Let $\rho = \frac{\psi'}{\psi}$, so that, from (2.1a),

$$\rho'(s) \geq A(s)^2 - 1 - \rho^2.$$

In case (i) we have $\rho' \geq \frac{A(s)^2}{2}$ on $[K_1, M]$ as long as $-1 \leq \rho < 0$, so that $\rho = 0$ (and hence $\psi' = 0$) before

$$\int_{K_1}^{t} \frac{A(s)^2}{2} ds = 1.$$

In other words,

$$\int_{K_1}^{M} \frac{A(s)^2}{2} ds \leq 1. \tag{2.12}$$

Thus, (2.10) implies that if $M \geq t \geq K_1$, then $\frac{A(s)}{2} \geq 1$ on $[t, M]$, and hence

$$\int_t^M \psi(s)^2 A(s) ds \leq \psi(K_1)^2 \int_t^M \frac{A(s)^2}{2} ds \leq \psi(K_1)^2.$$

Since (2.10) and (2.12) also imply that $M - K_1 < 1$, it follows from (2.12) and (2.9) that in case (i),

$$\int_{K_1}^{M} F(\psi, A)(t) dt < \delta \tag{2.13}$$

for sufficiently small ϵ and β.

In case (ii), since $\psi'(M) = 0$, there must be a $T \in [K_1, M)$ such that $\rho \leq -1$ on $[K_1, T]$ and $\rho \geq -1$ on $[T, M]$.

First consider $\int_{K_1}^{T} F(\psi, A)(t)dt$. On $[K_1, T]$ we have $\psi' \leq -\psi$, so that

$$\psi(t) \leq \psi(K_1)e^{K_1 - t}. \tag{2.14}$$

We now estimate the term

$$\int_t^M \psi(s)^2 A(s)ds$$

in $F(\psi, A)$ for $K_1 \leq t \leq T$. For this we use the following lemma.

LEMMA 2.8. $\lim_{t\to\infty} \psi_0(t)A^0(t) = 0$.

Proof. In our estimates of ψ_0 we saw that $\frac{\psi_0'}{\psi_0} + H_0(t) \to 0$. Since A^0 grows linearly, the result follows.

We now continue with the proof of Lemma 2.7. We choose K_1 so that in addition to the earlier constraints, $\psi_0(K_1)A^0(K_1) < 1$. For small ϵ and β, this inequality will also be satisfied by (ψ, A). If $t \leq T$, then

$$\begin{aligned}\int_t^M \psi(s)^2 A(s)ds &= \int_t^T \psi(s)^2 A(s)ds + \int_T^M \psi(s)^2 A(s)ds \\ &\leq \int_t^T \psi(s)^2 A(s)ds + \psi(T)^2 \int_T^M A(s)ds.\end{aligned}$$

The argument used earlier to get (2.12) shows that $\int_T^M A(s)ds \leq 1$, where we use (2.10) with A substituted for A^0. From an integration of $\frac{\psi'}{\psi} \leq -1$ we obtain

$$\int_t^T \psi(s)^2 A(s)ds \leq \psi(t)(\max_{t\leq s\leq T} \psi(s)A(s)) \int_t^T e^{t-s}ds.$$

In $[K_1, T]$, $\psi' \leq -\psi$, so $\frac{d}{ds}(\psi(s)A(s)) \leq -\psi(s)A(s) + \psi(s)A'(s)$. But $A'(s) \leq A'(M) = h \leq 2h^0$ for small ϵ and β, so $\psi(s)A(s)$ is decreasing in $[K_1, T]$. Hence, $\int_t^T \psi(s)^2 A(s)ds \leq \psi(t) \leq \psi(K_1)e^{K_1 - t}$. Using this we find that on $[K_1, T]$,

$$F(\psi, A) \leq r\psi(K_1)^2 e^{2(K_1 - t)}.$$

Thus, we can further restrict K_1 to ensure that for small positive ϵ and β,

$$\int_{K_1}^T F(\psi, A)(t)dt < \delta.$$

Finally, we consider $\int_T^M F(\psi, A)(t)dt$. As in case (i), we have

$$\int_T^M \frac{A(s)^2}{2}ds \leq 1.$$

Hence, we find that

$$\int_t^M \psi(s)^2 A(s)ds \leq \psi(T)^2 \leq \psi(K_1)^2$$

and

$$M - T \leq 1.$$

We then have (without further change in K_1) that

$$\int_T^M F(\psi, A)(t)dt < \delta.$$

Since δ was arbitrary, this completes the proof of Lemma 2.7.

3. Proof of Theorem 3. In order to prove Theorem 3 we need to show that the asymmetric solutions found in Theorem 2 have the additional property that their corresponding energy $\mathcal{G}$ is negative if $\epsilon > 0$ and $\beta > 0$ are small enough. Thus, we define

$$Q = \frac{2(m+M)}{\beta^4}\mathcal{G}. \tag{3.1}$$

Then, by (3.1) and (1.10), Q is given by

$$Q = \frac{1}{\beta^2}\int_{-m}^{M}\left(\psi^2\left(\frac{\beta^2\psi^2}{2} - 1 + A^2\right) + (\psi')^2 + \frac{1}{r\beta^2}(A' - h)^2\right)dt. \tag{3.2}$$

Therefore, if we show that $Q < 0$ for small $\epsilon > 0$ and $\beta > 0$ then $\mathcal{G}$ is also negative and Theorem 3 is proved. We need to simplify Q. For this we use (1.11) and conclude that

$$\psi^2\left(\frac{\beta^2\psi^2}{2} - 1 + A^2\right) + (\psi')^2 = \psi\psi'' + (\psi')^2 - \frac{\beta^2\psi^4}{2}. \tag{3.3}$$

Then substitution of (3.3) ino (3.2), together with the observation that $(\psi\psi')' = \psi\psi'' + (\psi')^2$, reduces Q to

$$Q = \frac{1}{\beta^2}\int_{-m}^{M}\left((\psi\psi')' - \frac{\beta^2\psi^4}{2} + \frac{1}{r\beta^2}(A' - h)^2\right)dt. \tag{3.4}$$

Since $\psi'(-m) = \psi'(M) = 0$, (3.4) further reduces to

$$Q = \frac{1}{\beta^2}\int_{-m}^{M}\left(\frac{1}{r\beta^2}(A' - h)^2 - \frac{\beta^2\psi^4}{2}\right)dt. \tag{3.5}$$

Next, it follows from an integration of (1.12) that

$$h - A' = r\beta^2\int_t^M \psi^2 A ds. \tag{3.6}$$

Finally, we substitute (3.6) into (3.5) and arrive at

$$Q = \int_{-m}^{M}\left(r\left(\int_t^M \psi^2 A ds\right)^2 - \frac{\psi^4}{2}\right)dt. \tag{3.7}$$

In view of Lemma 2.7 we conclude from (3.7) that

$$\lim_{\epsilon\to 0}\lim_{\beta\to 0} Q = \int_0^\infty \left(r\left(\int_t^\infty \psi_0^2 A^0 ds \right)^2 - \frac{\psi_0^4}{2} \right) dt, \tag{3.8}$$

where $A^0 = -1 + h^0 s$.

In the next section we prove the following lemma.

LEMMA 3.1. *There is a value $\gamma > 0$ such that*

$$\int_0^\infty \left(r\Big(\int_t^\infty \psi_0^2 A^0 ds \Big)^2 - \frac{\psi_0^4}{2} \right) dt < 0 \tag{3.9}$$

for all $r \in (0, 2+\gamma)$.

From Lemma 3.1 we observe that if $r \in (0, 2+\gamma)$, then $Q < 0$ for small $\epsilon > 0$ and $\beta > 0$. Therefore, $\mathcal{G}$ is also negative for small $\beta > 0$ and Theorem 3 is proved.

4. Proof of Lemma 3.1. For the proof of Lemma 3.1 we recall that ψ_0 satisfies

$$\psi_0'' = \psi_0(-2h^0 t + (h^0)^2 t^2), \tag{4.1}$$

$$\psi_0(0) = 1, \psi_0'(0) = 0, \tag{4.2}$$

where $h^0 > 0$ is the unique positive value for which ψ_0 satisfies $\psi_0' < 0$ for all $t > 0$, and

$$\lim_{t\to\infty} (\psi_0(t), \psi_0'(t)) = (0, 0). \tag{4.3}$$

The existence and uniqueness of h^0 were proved in Lemma 2.1. Thus, our main objective is to prove the following lemma.

LEMMA 4.1. *There is a value $\gamma > 0$ such that*

$$\int_0^\infty \left(r\Big(\int_t^\infty \psi_0^2(-1 + h^0 s) ds \Big)^2 - \frac{1}{2}\psi_0^4(t) \right) dt < 0 \tag{4.4}$$

for all $r \in (0, 2+\gamma)$.

The proof of Lemma 4.1 relies on an auxiliary result which we now establish.

Define the Riccati variable $q = \frac{\psi'(t)}{\psi(t)}$. Then q satisfies

$$q' + q^2 + 2h^0 t - (h^0)^2 t^2 = 0, \tag{4.5}$$

$$q(0) = q'(0) = 0. \tag{4.6}$$

LEMMA 4.2. *It follows from* (4.5) *and the definition of q that*

$$q < 0 \quad \textit{and} \quad q' < 0 \quad \textit{for all } t > 0, \tag{4.7}$$

$$q \le -\sqrt{(h^0)^2 t^2 - 2h^0 t} \quad \textit{for all } t \ge \frac{2}{h^0}, \tag{4.8}$$

$$q' > -h^0 \quad \textit{for all } t \ge 0. \tag{4.9}$$

Proof. Setting $h = h^0$ and $s = t\,h^0$, we observe that (4.7) and (4.8) follow immediately from Lemma 2.2. It remains to prove (4.9). It follows from (4.6) that $q'(0) = 0$. Thus $q' > -h^0$ on an interval $[0, \eta)$ for small $\eta > 0$. Suppose that (4.9) is false. Then there is a first $\hat{t} > 0$ for which $q'(\hat{t}) = -h^0$ and $q''(\hat{t}) \leq 0$. Two differentiations of (4.5) lead to

$$q''' + 2qq'' = 2((h^0)^2 - (q')^2). \tag{4.10}$$

One solution of (4.10) is $q' \equiv -h^0$ for all t. Thus, if $q''(\hat{t}) = 0$ then uniqueness of solutions implies that $q' = -h^0$ for all $t \geq 0$, and in particular, $q'(0) = -h^0$, contradicting the fact that $q'(0) = 0$. Therefore, it must be the case that

$$q'(\hat{t}) = h^0 \quad \text{and} \quad q''(\hat{t}) < 0. \tag{4.11}$$

It then follows from (4.10) and (4.11) that $q''(t) < 0$ for all $t > \hat{t}$ so that

$$\lim_{t\to\infty} q'(t) < -h^0. \tag{4.12}$$

We conclude from (4.12) that

$$\lim_{t\to\infty} q(t) + h^0 t < 0. \tag{4.13}$$

However, it follows from Lemma 2.2 that $\lim_{t\to\infty} q(t) + h^0 t = 0$, contradicting (4.13). Thus, it must be the case that $q' > -h^0$ for all $t > 0$ and the proof is complete.

We now return to the proof of Lemma 4.1. First, we conclude from (4.7) and (4.8), and the properties that $0 < \psi_0 < 1$ and $\psi_0' < 0$ for all $t > 0$, that the integral $\int_0^\infty \psi_0^4(t)dt$ is well defined and positive. The same reasoning shows that

$$J = \int_0^\infty \left(\int_t^\infty \psi_0^2(s)(-1 + h^0 s)ds \right)^2 dt \tag{4.14}$$

is well defined and positive. It remains to prove (4.4). For this we begin by estimating the integral

$$H = \int_t^\infty \psi_0^2(-1 + h^0 s)ds. \tag{4.15}$$

An integration by parts reduces (4.15) to

$$H = -\frac{\psi_0^2}{2h^0}(-1 + h^0 t)^2 - \frac{1}{h^0}\int_t^\infty \psi_0'\psi_0(-1 + h^0 s)^2 ds. \tag{4.16}$$

It follows from (4.1) that $\psi_0'\psi_0'' + \psi_0'\psi_0 = \psi_0'\psi_0(-1 + h^0 t)^2$, and therefore, (4.16) becomes

$$H = -\frac{\psi_0^2}{2h^0}(-1 + h^0 t)^2 + \frac{\psi_0^2(t)}{2h^0} + \frac{1}{2h^0}(\psi_0')^2. \tag{4.17}$$

Recall that $\psi' = q\psi$. Then (4.6) and (4.17) imply that

$$H = -\frac{\psi_0^2 q'}{2h^0}. \tag{4.18}$$

It follows from (4.18) and (4.14) that

$$J = \int_0^\infty H^2(t)dt \le \int_0^\infty \frac{\psi_0^4(t)}{4(h^0)^2}(q')^2 dt. \tag{4.19}$$

Because Lemma 4.2 gives $-h^0 < q' < 0$ for all $t > 0$, (4.19) further reduces to

$$J < \frac{1}{4}\int_0^\infty \psi_0^4(t)dt. \tag{4.20}$$

Finally, it follows from (4.4), (4.14), and (4.20) that there exists $\gamma > 0$ such that

$$\int_0^\infty \left(r\left(\int_t^\infty \psi_0^2(-1+h^0 s)ds \right)^2 - \frac{\psi_0^4}{2} \right) dt < 0$$

for all $r \in [0, 2+\gamma)$. This completes the proof of Lemma 4.1.

Appendix. In [9], when studying symmetric solutions, we used a different scaling from that used in this paper. That is, we rescaled (1.2)–(1.5) by setting

$$K = \frac{k^2 d^2}{4}, \quad h = \frac{h_e d}{2}, \quad r = \frac{1}{k^2}, \quad y = \frac{2x}{d},$$

and defined new dependent variables ϕ and a by

$$\phi(y) = \tilde{\phi}(x), \qquad a(y) = \tilde{a}(x).$$

This transforms the problem (1.2)–(1.5) into

$$\phi'' = K\phi(\phi^2 + a^2 - 1),$$

$$a'' = rK\phi^2 a,$$

$$\frac{d\phi}{dy}(\pm 1) = 0, \qquad \frac{da}{dy}(\pm 1) = h,$$

$$\phi > 0 \text{ on } [-1, 1].$$

Since we are considering symmetric solutions, we consider the initial values

$$\phi(0) = \beta, \;\; \phi'(0) = 0, \;\; a(0) = 0, \;\; a'(0) = \alpha.$$

In [11], Kwong proved that for each $\beta \in (0,1)$ there exists a unique $\alpha = \alpha(\beta) > 0$, continuously dependent and decreasing in β such that $\phi'(1, \beta, \alpha(\beta)) = 0$. He then set $h(\beta) = a'(1, \beta, \alpha(\beta))$ and proved that h is continuous, $h(0) > 0$, and $h(1) = 0$. Thus, to obtain the upper bound $h_{sym} \le \sqrt{3}$ given in [9] we found that it was sufficient to fix $r \in (0, 2.01]$ and estimate

$$h_{sym} = \overline{\lim}_{K\to\infty}\Big(\sup_{0\le\beta\le 1} \frac{h(\beta)}{\sqrt{K}} \Big).$$

In [9], our estimate for h_{sym} was obtained by carefully analyzing the behavior of (ϕ, a) for each $\beta \in (0,1)$. We considered three intervals of β, namely, $(0,.1], (.1, 1-\frac{1}{\sqrt{K}}]$, and $(1-\frac{1}{\sqrt{K}}, 1)$. We made repeated use of the "energy" function

$$\frac{\phi'^2}{K} + \frac{a'^2}{rK} - \frac{\phi^4}{2} + \phi^2 - a^2\phi^2,$$

which is constant for solutions of the system. We defined the function

$$Q(y) = \beta^2 - \frac{\beta^4}{2} + \frac{\phi(y)^4}{2} - \phi(y)^2 + a^2\phi(y)^2$$

and found the point y_0 where $Q(y_0) = 1$ to be important. The difficult part of the estimate required us to consider small intervals of values of $\phi(y_0)$, say, $I_1 < \phi(y_0) < I_2$. A priori, we know only that $0 < \phi(y_0) < 1$ and to get better estimates we had to subdivide $(0,1)$ into nine small subintervals $[I_1, I_2]$.

In this paper, to get the improved estimates required to show that $h_{sym} \le 1.68$ we have to use more subintervals. Thirty-four subintervals suffice. They are [0,.407], [.407,.56], [.56,.6], then 30 intervals from .6 to .9 in steps of .01, and finally [.9,1.0].

With this change, the proof in [9] gives the required upper bound for h_{sym}, and because of our rescaling this immediately leads to $h_e^{sym} \le 1.68k$ for $k \ge \frac{1}{\sqrt{2.01}}$. As we mentioned in [9], our estimate for h_{sym} is much easier for small r.

REFERENCES

[1] A. Aftalion, *On the minimizers of the Ginzburg–Landau energy for high kappa: The one-dimensional case*, European J. Appl. Math., 8 (1997), pp. 331–345.

[2] C. Bolley and B. Helffer, *Rigorous results for the G.L. Equations Ssociated to a Superconducting Film in the Weak κ Limit* preprint, Ecole Centrale de Nantes, Nantes, France, 1984.

[3] C. Bolley, *Modélisation du champ de retard à la condensation d'un supraconducteur par un probléme de bifurcation*, RAIRO Modél. Math. Anal. Numér., 26 (1992), pp. 175–287.

[4] C. Bolley and B. Helffer, *An application of semi-classical analysis to the asymptotic study of the supercooling field of a superconducting material*, Ann. Inst. H. Poincaré, 58 (1993), pp. 189–233.

[5] C. Bolley and B. Helffer, *Rigorous results on G.L. models in a film submitted to an exterior parallel magnetic field* preprint, Ecole Centrale de Nantes, Nantes, France, 1993.

[6] S. J. Chapman, *Nucleation of superconductivity in decreasing fields* I, European J. Appl. Math, 5 (1994), pp. 449–468.

[7] S. J. Chapman, S. D. Howison, and J. R. Ockendon, *Macroscopic models for superconductivity*, SIAM Rev., 34 (1992), pp. 529–560.

[8] V. L. Ginzburg and L. D. Landau, *On the theory of superconductors*, Soviet Phys. JETP, 20 (1950), p. 1064.

[9] S. P. Hastings and W. C. Troy, *On the existence of asymmetric minimizers for the one-dimensional Ginzburg–Landau model of superconductivity*, in Nonlinear Problems in Applied Mathematics, T. S. Angell, L. P. Cook, R. E. Kleinman, and W. E. Olmstead, eds., SIAM, Philadelphia, PA, 1996, pp. 150–158.

[10] S. P. Hastings, M. K. Kwong, and W. C. Troy, *The existence of multiple solutions for a Ginzburg–Landau type model of superconductivity*, European J. Appl. Math., 7 (1996), pp. 559–574.

[11] M. K. Kwong, *On the one-dimensional Ginzburg–Landau BVPs*, J. Differential Integral Equations, 8 (1995), pp. 1395–1405.

[12] E. M. Lieb and M. Loss, *Symmetry of the Ginzburg–Landau minimizer in a disc*, in Journees "Equations aux Derivees Patrielles" (Saint-Jean-de Monts, 1995), Exp. No. XVIII, Ecole Polytechnique, Palaiseau, France, 1995.

[13] P. Marcus, *Exact solutions of the Ginzburg-Landau equations for slabs in tangential magnetic fields*, Rev. Mod. Phys., 36 (1964), pp. 294–299.

[14] J. B. McLeod, *Ginzburg–Landau vortices,* in International Mathematics Conference '94 (Kaohsiung, 1994), World Sci. Publishing, River Edge, NJ, 1996, pp. 153–159.
[15] F. Odeh, *Existence and bifurcation theorems for the Ginzburg–Landau equations*, J. Math. Phys., 8 (1967), p. 4351.
[16] R. Seydel, *From Equilibrium to Chaos; Practical Bifurcation and Stability Analysis*, Elsevier, New York, 1988.
[17] R. Seydel, *Branch switching in bifurcation problems in ordinary differential equations*, Numer. Math., 41 (1983), pp. 93–116.
[18] S. Wang and Y. Yang, *Symmetric superconducting states in thin films*, SIAM J. Appl. Math., 52 (1992), pp. 614–629.
[19] Y. Yang, *Boundary value problems of the Ginzburg–Landau equations*, Proc. Roy. Soc. Edinburgh., 114A (1990), pp. 355–365.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 19–37

REGULARITY AND CONVERGENCE OF CRYSTALLINE MOTION*

KATSUYUKI ISHII† AND HALIL METE SONER‡

Abstract. We consider the motion of polygons by crystalline curvature. We show that "smooth" polygon evolves by crystalline curvature "smoothly" and that it shrinks to a point in finite time. We also establish the convergence of crystalline motion to the motion by mean curvature.

Key words. crystalline motion, motion by mean curvature, viscosity solutions

PII. S0036141097317347

1. Introduction.

Several models in phase transitions give rise to geometric equations relating the normal velocity of the interface to its curvature. The curvature term is related to surface tension and the surface energy is often an anisotropic function of the normal direction, indicating the preferred directions of the underlying crystal structure.

When the surface energy is isotropic, the resulting equation is the mean curvature flow and a variety of techniques have been used to analyze this flow. Huisken [25] showed that any convex set in higher than two space dimensions, shrinks to a point smoothly in finite time. We note that Huisken's method cannot be applied to the planar motion by mean curvature. Using different methods from those in [25], Gage and Hamilton [15] and Grayson [24] showed that a smooth planar embedded curve first becomes convex and then smoothly shrinks to a point in finite time. However, in general, in dimensions higher than two, embedded hypersurfaces may develop singularities and a weak formulation of the mean curvature flow is necessary to define the subsequent evolution after the onset of singularities. Brakke [8] was the first to study the mean curvature flow past the singularities. Using varifolds in geometric measure theory, he constructed global generalized solutions that are not necessarily unique. Almgren, Taylor, and Wang [2] used a time-step energy minimization approach together with geometric measure theory to analyze a very general class of equations.

An alternate approach, initially suggested in the physics literature by Ohta, Jasnaw, and Kawasaki [28], for numerical calculations by Osher and Sethian [26], represents the evolving surfaces as the level set of an auxiliary function solving an appropriate nonlinear differential equation. This level-set approach has been extensively developed by Chen, Giga, and Goto [9] and Evans and Spruck [12]. Evolution of hypersurfaces with codimension greater than one is studied by Ambrosio and Soner [3], and intrinsic definitions were developed by Soner [29] and Barles, Soner, and Souganidis [7]. Since the level-set equations are degenerate parabolic, the theory of viscosity solutions by Crandall and Lions [11] is used to define the level-set solutions. For more

*Received by the editors February 21, 1997; accepted for publication (in revised form) November 5, 1997; published electronically September 25, 1998.
http://www.siam.org/journals/sima/30-1/31734.html

†Department of Mathematics, Kobe University of Mercantile Marine, Higashinada, Kobe 658, Japan (ishii@cc.kshosen.ac.jp). This work was done while this author was visiting the Department of Mathematics, Carnegie Mellon University.

‡Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213. (mete@fermet.math.cmu.edu). This author was partially supported by the Army Research Office and the National Science Foundation through the Center for Nonlinear Analysis and by NSF grants DMS-9200801 and DMS-9500940 and by ARO grant DAAH04-95-1-0226.

information on viscosity solutions see the survey by Crandall, Ishii, and Lions [10] and the book by Fleming and Soner [13].

When the surface energy is convex, the evolution law is still degenerate parabolic and much of the above theory generalizes to these equations as well.

Nonsmooth energies are also of interest, and an interesting class of surface energies—called *crystalline* energies—have polygonal Frank diagrams. For these energies, the corresponding solutions are also polygonal, and the evolution law is a system of ordinary differential equations for the length of each side of the solution (see (2.3) below). An excellent introduction to crystalline motion is given in the recent book of Gurtin [22] and in the surveys of Taylor [32] and Taylor, Cahn, and Handwerker [34]. Short time existence and the other properties of the planar solutions are proved by Angenent and Gurtin [4] and Taylor [33]. Almgren and Taylor [1] showed that the crystalline flow is consistent with the variational approach developed in [2]. In a recent preprint Giga, Gurtin, and Mathias [19] study the classical solutions in three space dimensions and a deep viscosity theory for graph-like solutions of very general geometric equations have been developed by Giga and Giga [16] and the references therein. We also refer to Gurtin, Soner, and Souganidis [23] and Ohnuma and Sato [27], which treat a relaxed formulation of evolving surfaces by nonconvex interfacial energies.

In this paper, we consider a two-dimensional problem with a crystalline energy whose level sets are regular n-polygons and show the convergence of these solutions to the unique smooth solution of the mean curvature flow. This convergence has already been proved by Girao [20] for convex solutions and by Girao and Kohn [21] for graph-like solutions. They also obtained the rate of convergence. Here we generalize the convergence results in [20, 21] to general curves that are not necessarily convex. Our proof is a set theoretic analogue of the weak viscosity approach of Barles and Perthame [5, 6]. To describe our approach, let $\{\Omega_n(t)\}_{t\in[0,T)}$ be a sequence of open polygons each solving a crystalline flow. We define two possible limits:

$$\widehat{\Omega}(t) := \limsup_{n\to\infty,\ s\to t} \Omega_n(s),$$
$$\underline{\Omega}(t) := \liminf_{n\to\infty,\ s\to t} \Omega_n(s).$$

(Precise definitions are given in (4.2) below.) Then, with *only* L^∞ *estimates*, the Barles–Perthame approach enables us to show that $\widehat{\Omega}$ is a viscosity subsolution of the mean curvature flow, and $\underline{\Omega}$ is a viscosity supersolution of the mean curvature flow. Since, in two space dimensions, there is a smooth solution to the mean curvature flow, we show that both of these sets are equal to the smooth solution. This yields the convergence of Ω_n in the Hausdorff topology.

The paper is organized as follows. In the next section, we give the definition of crystalline motion and prove the existence of a regular solution in section 3. We define the weak viscosity limits in section 4 and prove their viscosity properties. Convergence is proved in the final section. Some properties of the viscosity solutions are gathered in the appendix.

After this work was completed, we were informed of a recent work of Giga and Giga [17] related to ours. They proved the stability of the periodic graph-like solutions for the motion by nonlocal weighted curvature. They also proved the motion by crystalline energy is shown to approximate the motion by regular interfacial energy if the crystalline energy approximates the regular interfacial energy. We also refer to Fukui and Giga [14] for an approximation property of the motion by nonsmooth weighted energy.

2. Crystalline motion and n-smooth polygons. Here we recall several standard definitions and equations. Gurtin's book [22] provides an excellent introduction to this subject. Also, see [31, 33].

2.1. Surface energy. All geometric flows that we consider are, formally, the gradient flows of the surface energy functional

$$I(\Gamma) := \int_\Gamma f(\vec{n})\, ds, \tag{2.1}$$

where Γ is a Jordan curve in $\mathcal{R}^2$, $\vec{n}$ is its outward unit normal vector, and $f : S^1 \to [0, \infty)$ is the *surface energy* function. It is customary to extend f to the whole $\mathcal{R}^2$ as a homogeneous function of degree one,

$$f(x) = |x| f\left(\frac{x}{|x|}\right) \qquad \forall x \neq 0,$$

and define

$$\hat{f}(\theta) := f(\cos\theta, \sin\theta).$$

Then the twice differentiability of f on $\mathcal{R}^2\backslash\{0\}$ is equivalent to the twice differentiability of $\hat{f}$, and f is convex if and only if $\hat{f}(\theta) + \hat{f}_{\theta\,\theta}(\theta) \geq 0$ for all θ.

The *Frank diagram* of the surface energy f is simply the polar graph of $\hat{f}^{-1}$, or equivalently, it is the one-level set of f, i.e.,

$$\mathcal{F}(f) := \{x \in \mathcal{R}^2 \ : \ f(x) = 1\} = \{r\,(\cos\theta, \sin\theta) \ : \ r\hat{f}(\theta) = 1\}.$$

When the surface tension f is smooth and convex, the gradient flow for the functional I has the form

$$\beta(\theta)V = (\hat{f}(\theta) + \hat{f}_{\theta\,\theta}(\theta))\ \kappa, \tag{2.2}$$

where V, κ, $(\cos\theta, \sin\theta)$ are, respectively, the normal velocity, the curvature, and the normal vector of the solution $\Gamma(t)$, and the given nonnegative function β is the kinetic coefficient. The mean curvature flow corresponds to $\hat{f} \equiv \beta \equiv 1$, and the other cases with strictly convex surface energy are qualitatively very similar to the mean curvature flow.

If f is not convex, we need to modify *both* f and β to obtain the correct relaxed equation. This relaxation procedure and the analytical properties of the relaxed equation was studied by Gurtin, Soner, and Souganidis [23] and, independently, by Ohnuma and Sato [27]. The common critical hypothesis in these works is the continuous differentiability of the relaxed surface energy function.

2.2. Crystalline flow. Nonsmooth energy functions are of interest in models for crystal growth, as it is well known that solid crystals can exist in polygonal shapes. An interesting class of nonsmooth energies are the *crystalline* energies. The Frank diagram of crystalline energy is a polygon.

Although the crystalline energies are only Lipschitz continuous, an appropriate weak formulation of (2.2) is possible and is called the crystalline flow; see [22, section 12.5] for the precise definition. The crystalline flow was derived by Taylor [31] and, independently, from thermodynamical considerations by Angenent and Gurtin [4].

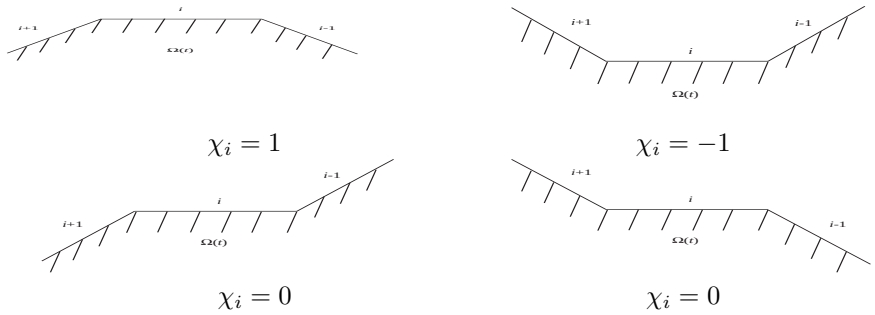


FIG. 1. *Definition of* χ_i.

Consider a crytalline energy function f, and let $\Theta := \{\theta_1, \dots, \theta_N\}$ be the angles corresponding to the corner points of the Frank digram of f. Suppose that the curve Γ is locally smooth around a point with a normal angle $\theta^* \notin \Theta$—say, $\theta^* \in (\theta_1, \theta_2)$. We can, then, decrease the energy $I(\Gamma)$ of Γ by infinitesimally alternating the normal angle between θ_1 and θ_2. Therefore, for crystalline energies, we consider only polygonal solutions with normal angles taking values in Θ.

In this paper, for simplicity, we consider only crystalline energies whose Frank diagrams are regular n-polygons, and kinetic coefficient $\beta \equiv 1$. Then

$$\Theta = \Theta_n := \left\{ \frac{2\pi k}{n} \ : \ k = 0, 1, \dots, (n-1) \right\}.$$

Here and hereafter $\theta \in \Theta$ means $\theta \equiv 2\pi k/n \mod 2\pi$ for some $k \in \{0, 1, \dots, n-1\}$. The evolution of side i, $L_i(t)$, is governed by

$$V_i(t) = -\frac{2\tan(\pi/n)}{l_i(t)}\ \chi_i, \tag{2.3}$$

where $V_i(t)$, $l_i(t)$, and χ_i, are, respectively, the normal velocity, the length, and the discrete curvature of $L_i(t)$. The discrete curvature $\chi_i \in \{-1, 0, +1\}$. It is equal to $+1$ if both edges of $L_i(t)$ have positive curvature, it is equal to -1 if both edges of $L_i(t)$ have negative curvature, and it is equal to zero otherwise; see Figure 1. ($\Omega(t)$ denotes the domain enclosed by $L_i(t)'s$.)

We close this subsection by stating the evolution rule for the length, $l_i(t)$, of the sides of a solution of the crystalline flow:

$$\frac{d}{dt} l_i(t) = \frac{1}{\cos^2(\pi/n)} \left(2\cos\left(\frac{2\pi}{n}\right) \cdot \frac{\chi_i^2}{l_i(t)} - \frac{\chi_{i+1}^2}{l_{i+1}(t)} - \frac{\chi_{i-1}^2}{l_{i-1}(t)} \right). \tag{2.4}$$

This equation follows from (2.3) and geometry; see [22, equation (12.39)].

2.3. *n*-smooth polygons. We continue by defining the notion of a "good" solution of (2.3). For a polygon Γ, let $N(\Gamma)$ be the total number of sides.

DEFINITION 2.1. *We say that a closed polygon* Γ *is an* n*-smooth polygon if* $N(\Gamma)$ *is finite and*

(1) Γ *encloses a simply-connected, bounded, open subset of* $\mathcal{R}^2$,
(2) *for every* $i = 1, \dots, N(\Gamma)$, *the normal angle* θ_i *of the side* i *belongs to* Θ_n,
(3) $|\theta_i - \theta_{i-1}| = 2\pi/n$ *for every* $i = 1, \dots, N(\Gamma)$, *where* $|\theta_i - \theta_{i-1}|$ *is understood as the infimum over its representatives.*

The third condition is formally equivalent to the "discrete continuity" of the normal angle, which explains the term "smooth."

By definition, any solution of (2.3) satisfies the second condition.

Let

$$N^+(\Gamma) := \{i \in \{1, \dots, N(\Gamma)\} \ : \chi_i = 1\},$$

$$N^-(\Gamma) := \{i \in \{1, \dots, N(\Gamma)\} \ : \chi_i = -1\},$$

$$N^0(\Gamma) := \{i \in \{1, \dots, N(\Gamma)\} \ : \chi_i = 0\}.$$

Then for any n-smooth polygon Γ,

$$N^+(\Gamma) - N^-(\Gamma) = \sum_{i=1}^{N(\Gamma)} \chi_i = n \tag{2.5}$$

is an identity which is the discrete version of

$$\int_C \kappa \, ds = 2\pi$$

for a smooth Jordan curve C.

3. Regularity. In this section, we will show that there is a unique n-smooth solution of (2.3) which evolves smoothly in time (i.e., remains n-smooth) and shrinks to a point in finite time. This is the discrete analogue of a theorem of Grayson [24] and Gage and Hamilton [15]. A more general statement is proved by Taylor [33, Theorem 3.1]. For the reader's convenience, we provide all the details of this result.

THEOREM 3.1 (Taylor [33]). *Let Γ_0 be an n-smooth polygon enclosing an open set Ω_0. Then there exist n-smooth polygons $\{\Gamma(t)\}_{t\in[0,T)}$ solving (2.3) with the initial condition $\Gamma(0) = \Gamma_0$. Moreover $\Gamma(t)$ shrinks to a point as $t \uparrow T$, and*

$$T = \frac{|\Omega_0|}{2n \tan(\pi/n)}. \tag{3.1}$$

Remark 3.2. Uniqueness follows from Giga and Gurtin [18] and Taylor [33].

We start with several results toward the proof of Theorem 3.1.

Clearly, for a short time there is a solution $\Gamma(t)$ satisfying initial data. Let $t_1 > 0$ be the first time this solution is no longer n-smooth. Since, by definition, the normal angles of any solution take values in Θ_n (cf. section 2.2), there are two possibilities at t_1: either the length of one or more sides tend to zero or the solution self-intersects at t_1. We will first show that the latter does not happen. Our proof is very similar to [33, Theorem 3.2(1)].

LEMMA 3.3. *Let t_1 and $\{\Gamma(t) = \partial\Omega(t)\}_{t\in[0,t_1)}$ be as above. Then*

$$\liminf_{t\uparrow t_1} \inf\{l_i(s) \ : \ s \in [0,t], \ i = 1, \dots, N(\Gamma(0))\} = 0.$$

Proof. Suppose the opposite. Then

$$\inf\{l_i(s) \ : \ s \in [0,t_1), \ i = 1, \dots, N(\Gamma(0))\} > 0.$$

Then, by (2.4), each $l_i(\cdot)$ is smooth on $(0,t_1)$ and therefore

$$\Omega(t_1) = \lim_{t\uparrow t_1} \Omega(t)$$

exists in the Hausdorff topology. By the definition of t_1, $\Gamma(t_1)$ self-intersects. Moreover, for all $t \in [0,t_1]$,

$$|\theta_i - \theta_{i-1}| = \frac{2\pi}{n},\ i = 1,\dots,N(\Gamma(t)) = N(\Gamma(0)), \tag{3.2}$$

so that at t_1 there are two possibilities: either two sides or two corner points touch each other. Note that, by (3.2), if a corner point touches a side, then necessarily two sides also touch each other. The following arguments are very similar to those in [18].

Case 1. Suppose that $L_i(t_1)$ intersects at $L_j(t_1)$.

Then a straightforward analysis argument shows that $(\chi_i,\chi_j) = (1,-1)$ or $(\chi_i,\chi_j) = (-1,1)$. Since the analyses of these cases are symmetric, we may assume $(\chi_i,\chi_j) = (1,-1)$. Then $l_i(t_1) \le l_j(t_1)$.

Subcase (1). $l_i(t_1) < l_j(t_1)$.

Then for some $\delta > 0$, $l_i(t) < l_j(t)$ in $(t_1-\delta,t_1]$, and therefore,

$$\alpha(t) := \frac{2\tan(\pi/n)}{l_j(t)} - \frac{2\tan(\pi/n)}{l_i(t)} > 0, \qquad t \in (t_1-\delta,t_1].$$

But $\alpha(t)$ is equal to the time derivative of the distance between $L_i(t)$ and $L_j(t)$ and this distance is equal to zero at t_1. Hence this case is not possible.

Subcase (2). $l_i(t_1) = l_j(t_1)$.

Then, the sides adjacent to $L_i(t)$ and $L_j(t)$ also touch each other at time t_1, and therefore, there have to be two sides satisfying the assumptions of the previous subcase, thus yielding a contradiction.

Case 2. Two corner points touch each other.

Let the intersection, $x_i(t)$ of $L_i(t)$ and $L_{i+1}(t)$ be the same as the intersection $x_j(t)$ of the sides $L_{j-1}(t)$ and $L_j(t)$. Then the angle between $L_i(t)$ and $L_j(t)$ and the one between $L_{i+1}(t)$ and $L_{j-1}(t)$ are equal to $2\pi/n$. By rotation, we may assume that $L_i(t)$ and $L_j(t)$ are parallel to the x-axis, and $L_{i+1}(t)$ is aligned with the $L_{j-1}(t)$ (cf. Figure 2). Moreover, $\chi_k \ge 0$ for $k = i, i+1, j, j-1$. Let $V_{x_i}(t)$ and $V_{x_j}(t)$ be the velocity vectors of the points $x_i(t)$ and $x_j(t)$, respectively. Then

$$(0,1)\cdot(V_{x_j} - V_{x_i}) \ge 0,$$

and the inequality is strict unless $\chi_k = 0$ for all $k = i, i+1, j, j-1$. Since $x_i(t_1) = x_j(t_1)$, we conclude that $\chi_k = 0$ for all $k = i, i+1, j, j-1$. But then $V_{x_i}(t) = V_{x_j}(t) = 0$ for $t < t_1$ close to t_1 and this contradicts the definition of t_1. □

Our next result is the following lemma.

LEMMA 3.4. *Let t_1 and $\{\Gamma(t) = \partial\Omega(t)\}_{t\in[0,t)}$ be as above. Suppose t_1 is strictly less than the extinction time. Then as $t \to t_1$, $\Omega(t)$ converges to an n-smooth polygon $\Omega(t_1)$ in the Hausdorff topology.*

Proof. By the previous lemma, there is a side i^* such that

$$\liminf_{t\to t_1} l_{i^*}(t) = 0.$$

The main step in this proof is to show $\chi_{i^*} = 0$ if the side L_{i^*} disappears at t_1. So we suppose that it is equal to $+1$ or -1. Since the analyses of these cases are similar, we may assume that $\chi_{i^*} = 1$. Set $\theta = 2\pi/n$.

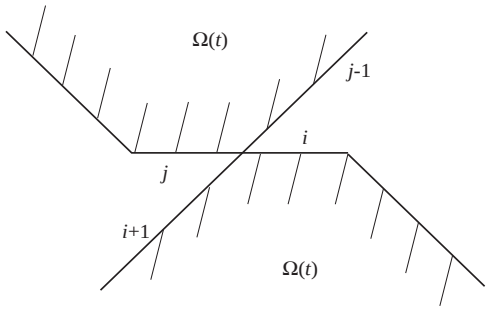


FIG. 2. *Case* 2.

1. In this step we will show that $l_{i^*}(\cdot)$ is continuous on $[0, t_1]$. For future reference, we will prove that, for any j, $l_j(\cdot)$ is continuous on $[0, t_1]$. By (2.4), all sides remain bounded, and we set

$$B := \limsup_{t \to t_1} l_j(t).$$

Suppose that

$$B > \liminf_{t \to t_1} l_j(t) := A.$$

Since $l_j(\cdot)$ is continuous in $[0, t_1)$, it crosses $(A + B)/2$ infinitely many times before t_1. In particular, by the mean value theorem, there is a sequence $t_k \uparrow t_1$ such that

$$l_j(t_k) \geq \frac{A+B}{2}, \quad \lim_{k \to +\infty} l_j'(t_k) = +\infty.$$

However, by (2.4),

$$l_j'(t_k) \leq \frac{2\cos\theta}{l_j(t_k)\cos^2(\theta/2)} \leq C$$

for some constant C independent of k. Hence $A = B$.

2. This step closely follows [33, Proposition 3.1].

Since t_1 is strictly less than the extinction time, there are at least two sides which have nonzero length at time t_1. Hence there are two sides L_{p_0} and L_{p_1} such that $p_0 < i^* < p_1$, $l_{p_0}(t)$ and $l_{p_1}(t)$ are uniformly positive in $[0, t_1]$, and

$$\lim_{t \uparrow t_1} l_j(t) = 0 \ \forall j = p_0 + 1, \ldots, p_1 - 1.$$

For any j, let $\mathcal{L}_j(t)$ be the line extending $L_j(t)$, $x_{j+1}(t)$ be the intersection between $\mathcal{L}_j(t)$ and $\mathcal{L}_{j+1}(t)$, and θ_j be the angle between the outward normal and the horizontal axis. Then, as $t \uparrow t_1$, all $x_{p_0+1}(t), \ldots, x_q(t)$ converge to the same point x^*.

We analyze several cases separately.

Case 1. $\chi_j \neq 0 \ \forall \ j = p_0 + 1, \ldots, p_1 - 1$.

Since we have assumed that $\chi_{i^*} = 1$, $\chi_j = 1 \ \forall \ j = p_0 + 1, \ldots, p_1 - 1$ and

$$x^* \in \bigcap_{0 \leq t < t_1} \bigcap_{j=p_0}^{p_1} \{y \in \mathcal{R}^2 \ : \ (y - x_j(t)) \cdot (\cos\theta_j, \sin\theta_j) \leq 0\}.$$

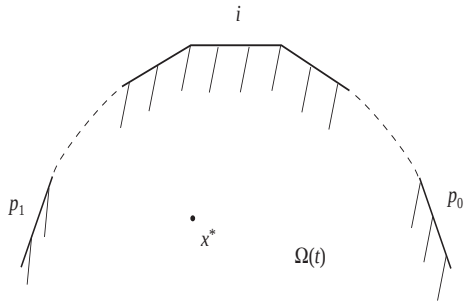


FIG. 3. *Position of* x^*.

See Figure 3.

By geometry, $|\theta_{p_0} - \theta_{p_1}| \leq \pi$.

Subcase 1. $|\theta_{p_0} - \theta_{p_1}| < \pi$.

Let $y(t)$ be the intersection between $\mathcal{L}_{p_0}(t)$ and $\mathcal{L}_{p_1}(t)$. We define

$$d(t) = (y(t) - x^*) \cdot (\cos\theta_{p_0+1}, \sin\theta_{p_0+1}),$$
$$d_{p_0+1}(t) = \operatorname{dist}(x^*, \mathcal{L}_{p_0+1}(t)).$$

Then $d_{p_0+1}(t) \leq d(t)\ \forall\ t \in [0, t_1)$ and $d_{p_0+1}(t_1) = d(t_1) = 0$. Moreover, $d(t)$ is Lipschitz continuous in t and

$$\frac{d}{dt} d_{p_0+1}(t) = V_{p_0+1}(t) = -\frac{2\tan(\theta/2)}{l_{p_0+1}(t)}.$$

Hence,

$$0 \geq -\int_t^{t_1} \frac{2\tan(\theta/2)}{l_{p_0+1}(\tau)} d\tau = d_{p_0+1}(t) \geq d(t) \geq -\|d'\|_{L^\infty(0,t_1)}(t_1 - t) \qquad \forall t < t_1.$$

This contradicts the fact $l_{p_0+1}(t) \to 0$ as $t \uparrow t_1$.

Subcase 2. $|\theta_{p_0} - \theta_{p_1}| = \pi$.

We repeat the argument used in the previous case with

$$\tilde{d}(t) = \operatorname{dist}(\mathcal{L}_{p_0}(t), \mathcal{L}_{p_1}(t)),$$
$$\tilde{d}_{p_0+1}(t) = \operatorname{dist}(L_{p_0+1}(t), \mathcal{L}_{p_1}(t)).$$

Case 2. $\chi_q = 0$ exactly for one $q \in \{p_0 + 1, \ldots, p_1 - 1\}$.

Then, $\chi_j = 1$ for $j = p_0 + 1, \ldots, q - 1$ and $\chi_j = -1$ for $j = q + 1, \ldots, p_1 - 1$, or $\chi_j = -1$ for $j = p_0 + 1, \ldots, q - 1$ and $\chi_j = 1$ for $j = q + 1, \ldots, p_1 - 1$. Since the arguments in both cases are similar, without loss of generality, we consider only the first possibility.

If $|\theta_{p_0} - \theta_q| \leq \pi$, we argue as in *Case* 1, using side $L_q(t)$ instead of $L_{p_1}(t)$. We also argue similarly, when $|\theta_q - \theta_{p_1}| \leq \pi$. Therefore, we may assume that $|\theta_{p_0} - \theta_q| > \pi$ and that there is a side $L_j(t)$ with $q < j < p_1$, which is parallel to $L_{p_0}(t)$. Let $\mathcal{L}$ be the line going through x^* and parallel to both $L_{p_0}(t)$ and $L_j(t)$. Set

$$d(t) = \operatorname{dist}(L_{p_0}(t), \mathcal{L}) - \operatorname{dist}(L_j(t), \mathcal{L}).$$

Then $0 = d(t_1)$ and since $|\theta_{p_0} - \theta_q| > \pi$, $0 < d(t)\ \forall\ (0, t_1)$; see Figure 4.

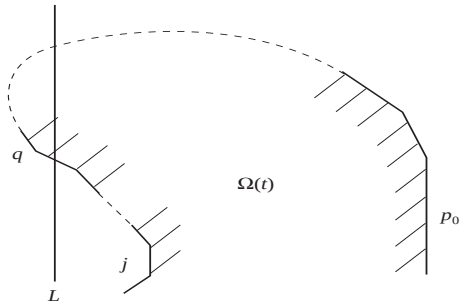


FIG. 4. *Case* 2.

However, this contradicts the fact that $d'(t) > 0 \ \forall\ t$ sufficiently close to t_1.

Case 3. $\chi_j = 0$ for more than one side.

Suppose that χ_q and χ_j are equal to zero. Then x^* belongs to both $L_q(t)$ and $L_j(t) \ \forall\ t$, and therefore, $j = q - 1$ or $q + 1$. Since $l_q(t)$ converges to zero, at least one side adjacent to $L_q(t)$ has nonzero discrete curvature. Hence there are two sides with zero discrete curvature and they are adjacent to each other. As in *Case* 1, all the other sides between $L_{p_0}(t)$ and $L_{p_1}(t)$ satisfy $\chi_k = 1$, and we argue as in *Case* 1.

Therefore, the case $\chi_{i^*} = 1$ is not possible. An entirely similar argument shows that the case $\chi_{i^*} = -1$ is not possible either. Hence $\chi_{i^*} = 0$ and L_{i^*-1} and L_{i^*+1} are parallel, and the normal angle of the "new" side is equal to that of these two ones. □

We are now in a position to prove Theorem 3.1.

Proof of Theorem 3.1. Since $\Gamma(0)$ is n-smooth for short time, there is an n-smooth solution $\Gamma(t)$. Moreover, by Lemma 3.4, this solution remains n-smooth until one side of $\Gamma(t)$ vanishes. Let t_1 be the first time a side vanishes. Then, $\Gamma(t)$ is n-smooth and $N(\Gamma(t)) = N(\Gamma(0)) \ \forall\ t \in [0, t_1)$. By Lemma 3.3, $\Gamma(t_1)$ is also n-smooth and $N(\Gamma(t_1)) \le N(\Gamma(0)) - 2$. We repeat this procedure starting from $\Gamma(t_1)$. Since $N(\Gamma(0))$ is finite, we have only to repeat finitely many times.

Let $t_1 < t_2 < \cdots < t_N$ be the times at which a side vanishes. Let $t_N > 0$ be the time when $N^-(\Gamma(t_N)) = N^0(\Gamma(t_N)) = 0$. Then, by (2.5), $N^+(\Gamma(t_N)) = n$ and $\Gamma(t)$ is convex for all $t \ge t_N$.

We see that $\Gamma(t)$ shrinks to a point at finite time. Indeed, by (2.5), we can calculate the rate of change of $|\Omega(t)|$:

$$\begin{aligned}\frac{d}{dt}|\Omega(t)| &= \sum_i V_i l_i \\ &= -\sum_{i \in N^+(\Gamma(t))} 2\tan\frac{\pi}{n} + \sum_{i \in N^-(\Gamma(t))} 2\tan\frac{\pi}{n} \\ &= -2n\tan\frac{\pi}{n}.\end{aligned}$$

From the foregoing calculation, we conclude that the solution shrinks to a point at some time T. Moreover, at time T,

$$0 = |\Omega(T)| = |\Omega_0| - 2n\tan\frac{\pi}{n}\cdot T,$$

and (3.1) follows. □

4. Weak viscosity limits. In this section, we will study the properties of the set-theoretic analogue of the weak viscosity limits of Barles and Perthame [5, 6]. Let $\{\Gamma_n(t)\}_{t\in[0,T)}$ be a sequence of n-smooth solutions of (2.3), and let $\Omega_n(t)$ be the open set enclosed by $\Gamma_n(t)$. Assume that there is a constant $R > 0$, independent of n, satisfying

$$\Omega_n(t) \subset B(0,R), \tag{4.1}$$

where $B(x,r) = \{y \in \mathcal{R}^2 \ : \ |y-x| \le r\}$. Following [6, 29], for $t \in [0,T)$, we define

$$\widehat{\Omega}(t) := \bigcap_{\substack{r>0\\N\ge 1}} \operatorname{cl}\left(\bigcup_{\substack{|s-t|\le r,\ 0\le s<T\\ n\ge N}} \Omega_n(s) \right), \tag{4.2}$$

$$\underline{\Omega}(t) := \bigcup_{\substack{r>0\\N\ge 1}} \operatorname{int}\left(\bigcap_{\substack{|s-t|\le r,\ 0\le s<T\\ n\ge N}} \Omega_n(s) \right),$$

where $\operatorname{cl} A$ and $\operatorname{int} A$ are, respectively, the closure and the interior of the set A. In view of (4.1), $\widehat{\Omega}(t)$ is a bounded closed set and $\underline{\Omega}(t)$ is a bounded open set. We will show that, respectively, $\widehat{\Omega}(t)$ is a weak subsolution and $\underline{\Omega}(t)$ is a weak supersolution of the mean curvature flow.

This type of stability results are typical in the theory of viscosity solutions and, in general, they are a simple consequence of the maximum principle. However, the crystalline flow is not defined for smooth curves and this fact is the major difficulty in the following analysis.

The notion of viscosity solutions we use is first introduced by the second author in [29] and further developed in [7, 30]. Here we only recall the definition; other relevant definitions and results are gathered in the appendix.

We continue by recalling several definitions that will be used in the subsequent analysis. For subsets $\{\Omega(t)\}_{0\le t<T}$ in $\mathcal{R}^2$, the *upper semicontinuous* (*u.s.c.*) *envelope* and, respectively, the *lower semicontinuous* (*(l.s.c.*) *envelope* are defined by

$$\Omega^*(t) = \bigcap_{r>0} \operatorname{cl}\left(\bigcup_{\substack{|s-t|\le r\\ 0\le s<T}} \Omega(s) \right), \quad \Omega_*(t) = \bigcup_{r>0} \operatorname{int}\left(\bigcap_{\substack{|s-t|\le r\\ 0\le s<T}} \Omega(s) \right), \quad t \in [0,T).$$

Then, it is clear that $(\underline{\Omega})_* = \underline{\Omega}$ and $(\widehat{\Omega})^* = \widehat{\Omega}$. For other properties of these envelopes, see [29, Lemma 3.1].

For a collection of closed subsets $\{O(t)\}_{0\le t<T}$ with smooth boundary, $V_O(x,t)$ is the normal velocity of $\partial O(t)$ at x and $\kappa_O(x,t)$ is the curvature of $\partial O(t)$ at x. We use the convention that the curvature of a convex curve is nonnegative.

We are now in a position to give the weak (viscosity) definition of the mean curvature flow we will use. This definition is very similar to the one given in [29]; see the appendix for the connection between these two definitions.

DEFINITION 4.1. *Let $\{\Omega(t)\}_{0\le t<T}$ be a collection of bounded subsets in $\mathcal{R}^2$ satisfying $\Omega_*(t) \neq \emptyset$ for every $t \in [0,T)$.*

We say $\{\Omega(t)\}_{0\le t<T}$ is a *weak subsolution* of the mean curvature flow, if for any closed, smooth subsets $\{O(t)\}_{0\le t<T}$,

$$V_O(x_0,t_0) \le -\kappa_O(x_0,t_0) \tag{4.3}$$

at each $t_0 \in (0,T)$ and $x_0 \in \partial O(t_0)$ satisfying

$$\Omega^*(t) \subset\subset O(t) \qquad \forall\, t \neq t_0, \tag{4.4}$$

$$\Omega^*(t_0) \subset O(t_0), \quad \partial\Omega^*(t_0) \cap \partial O(t_0) = \{x_0\}.$$

Similarly, we say $\{\Omega(t)\}_{0\le t<T}$ is a *weak supersolution* of the mean curvature flow if for any closed, smooth subsets $\{O(t)\}_{0\le t<T}$,

$$V_O(x_0,t_0) \ge -\kappa_O(x_0,t_0)$$

at each $t_0 \in (0,T)$ and $x_0 \in \partial O(t_0)$ satisfying

$$O(t) \subset\subset \Omega_*(t)\ \forall\, t \neq t_0, \quad O(t_0) \subset \Omega_*(t_0), \quad \partial\Omega_*(t_0) \cap \partial O(t_0) = \{x_0\}.$$

Condition (4.4) implies that $(x_0,t_0) \in \partial O(t_0) \times (0,T)$ is the *strict* maximizer of $-\mathrm{dist}(x, \partial\Omega^*(t))$ over all $(x,t) \in \partial O(t) \times (0,T)$. A similar conclusion also holds for supersolutions.

Following is the set theoretic analogue of the Barles and Perthame procedure [5, 6], [13, section 5], and it is the chief technical contribution of this paper.

Recall that $\Gamma_n(t) = \partial\Omega_n(t)$.

LEMMA 4.2. *$\widehat{\Omega}$ is a weak subsolution of the mean curvature flow, while $\underline{\Omega}$ is a weak supersolution.*

Before we give the proof of this lemma, we will first give a formal proof of the subsolution property.

In view of our definition of a weak solution, we start with smooth sets $\{O(t)\}_{0<t<T}$ and a point (x_0,t_0) satisfying (4.4). Our goal is to verify (4.3). By (4.4) there are a subsequence n_k and a sequence $(x_k,t_k) \to (x_0,t_0)$ satisfying $\Omega_{n_k}(t_k) \subset O(t_k)$ and that $x_k \in \Gamma_{n_k}(t_k)$. Although there are several other cases, assume that x_k is the intersection of $L_{i-1}(t_k)$ and $L_i(t_k)$ of $\Gamma_{n_k}(t_k)$, and $\chi_i = \chi_{i-1} = 1$. We choose a coordinate system so that x_k is the origin and the $L_i(t_k)$ side is included in the x_1-axis. Let $n_1 = (0,1)$, $n_2 = (\sin(2\pi/n_k), \cos(2\pi/n_k))$. Then, the unit normal vector of ∂O satisfies $n_O(x_k,t_k) = (\sin\alpha, \cos\alpha)$ for some $0 < \alpha < 2\pi/n_k$. By the crystalline equation (2.3),

$$V_{x_k} \cdot n_1 = V_i = -\frac{2\tan(\pi/n_k)}{l_i},$$
$$V_{x_k} \cdot n_2 = V_{i-1} = -\frac{2\tan(\pi/n_k)}{l_{i-1}},$$

and therefore,

$$V_{x_k} = 2\tan\frac{\pi}{n_k}\left(\frac{1}{\tan(2\pi/n_k)}\left(\frac{1}{l_i} - \frac{1}{l_{i-1}}\right), -\frac{1}{l_i}\right), \tag{4.5}$$

$$\begin{aligned} V_O(x_k,t_k) &= V_{x_k} \cdot n_O(x_k,t_k) \\ &= -\frac{1}{\cos^2(\pi/n_k)}\left(\frac{\sin(2\pi/n_k - \alpha)}{l_i} + \frac{\sin\alpha}{l_{i-1}}\right). \end{aligned} \tag{4.6}$$

Since $V_O(x_k,t_k)<0$, we may assume $\inf_{k\in\mathcal{N}}\kappa_O(x_k,t_k)>0$. This implies that, as $k\to\infty$, both l_i and l_{i-1} converge to zero. By elementary geometry, we obtain a sharper estimate: for every $\varepsilon>0$,

$$l_i\le\frac{2\sin\alpha}{\kappa_O(x_k,t_k)-\varepsilon},\qquad l_{i-1}\le\frac{2\sin(2\pi/n_k-\alpha)}{\kappa_O(x_k,t_k)-\varepsilon}$$

for sufficiently large k. Substitute these into (4.6):

$$\begin{aligned}V_O(x_k,t_k)&\le-\frac{\kappa_O(x_k,t_k)-\varepsilon}{2\cos^2(\pi/n_k)}\left(\frac{\sin(2\pi/n_k-\alpha)}{\sin\alpha}+\frac{\sin\alpha}{\sin(2\pi/n_k-\alpha)}\right)\\&\le-\kappa_O(x_k,t_k)+\varepsilon.\end{aligned}$$

In the foregoing argument, we crucially used the assumption that x_k is a "convex" corner point of Γ_{n_k}. Although this is the most likely situation, other cases may also arise, and for that we will perturb the test sets O in the preceding proof.

Proof. We will prove only the subsolution property. Proof of the supersolution case is similar.

Let $\{O(t)\}_{0<t<T}$ and (t_0,x_0) be as in (4.4). Our goal is to verify (4.3), i.e.,

$$v:=V_O(x_0,t_0)\le-\kappa:=-\kappa_O(x_0,t_0).$$

If necessary, by perturbing $O(\cdot)$, we may assume that $\kappa\neq0$. We analyze two cases separately.

Case 1. $\kappa>0$.

For $\varepsilon>0$, $x^*\in\mathcal{R}^2$, and a large constant K, let $D^\varepsilon(t:x^*)$ be the disk with center x^* and radius

$$R^\varepsilon(t)=\frac{1}{\kappa-\varepsilon}+v(t-t_0)+K(t-t_0)^2.$$

Set

$$x_0^\varepsilon:=x_0-R^\varepsilon(t_0)n_O(x_0,t_0).$$

By the smoothness of ∂O, for all sufficiently large K, there is a δ^ε such that

$$O(t)\cap B(x_0,2\delta^\varepsilon)\subset D^\varepsilon(t:x_0)\cap B(x_0,2\delta^\varepsilon)\tag{4.7}$$

for all $|t-t_0|\le2\delta^\varepsilon$. We fix K large enough so that the above inequality holds.

Next we approximate $D^\varepsilon(t:x^*)$ by regions with polygonal boundaries. Let

$$C_n:=\left\{\ x\in\mathcal{R}^2\ :\ x\cdot\left(\cos\left(\frac{2l\pi}{n}\right),\sin\left(\frac{2l\pi}{n}\right)\right)\le1\quad\forall\ l=0,1,\dots,(n-1)\ \right\},$$

and, for any x^*, set

$$D_n^\varepsilon(t:x^*):=\{x^*\}\oplus R^\varepsilon(t)C_n.$$

Since $D_n^\varepsilon(\cdot:x_0^\varepsilon)$ approximates $D^\varepsilon(\cdot:x_0^\varepsilon)$, by (4.4) and (4.7), there are a subsequence n_k and sequences $(x_k,t_k)\to(x_0,t_0)$, $y_k\to x_0^\varepsilon$ satisfying

$$x_k\in\Gamma_{n_k}(t_k)\cap\partial D_{n_k}^\varepsilon(t_k:y_k),$$

$$\Omega_{n_k}(t) \cap B(x_0, \delta^\varepsilon) \subset D^\varepsilon_{n_k}(t : y_k) \cap B(x_0, \delta^\varepsilon) \qquad \forall\ |t - t_0| \le \delta^\varepsilon.$$

A proof of this fact is given in the appendix in Lemma 6.2. To simplify the notations, we assume that $n_k = k$ and write $D_k(t)$ for $D^\varepsilon_{n_k}(t : y_k)$.

Let x_k be on the ith side of $\Gamma_k(t_k)$. Then the normal velocity, V_i, of this side is equal to the normal velocity of D_k at t_k. Hence,

$$V_i = v + 2K(t_k - t_0).$$

Since $D_k(t_k)$ is a regular k-polygon, $\chi_i(t_k) = 1$ and, therefore, the length $l_i(t_k)$ of side i of $\Gamma_k(t_k)$ is less than or equal to the length of any side of $D_k(t_k)$:

$$l_i(t_k) \le 2R^\varepsilon(t_k) \sin \frac{\pi}{k}.$$

Then, by (2.3) and the foregoing discussion,

$$v + 2K(t_k - t_0) = V_i = -\frac{2\tan(\pi/k)}{l_i(t_k)} \le -\frac{1}{R^\varepsilon(t_k)\cos(\pi/k)}.$$

Since $R^\varepsilon(t_k)$ converges to $1/\kappa$ and $t_k \to t_0$, we obtain (4.3) by first letting $k \to \infty$ and then $\varepsilon \downarrow 0$.

Case 2. $\kappa < 0$.

For small $\varepsilon > 0$ and any x^*, let $x_0^\varepsilon := x_0 + R^\varepsilon(t_0) n_O(x_0, t_0)$, and let $D^\varepsilon(t : x^*)$ be the complement of the disk with center x^*, radius

$$R^\varepsilon(t) = \frac{1}{-\kappa + \varepsilon} + v(t - t_0) - K(t - t_0)^2.$$

As in the previous case, there is a δ^ε such that

$$O(t) \cap B(x_0, 2\delta^\varepsilon) \subset D^\varepsilon(t : x_0^\varepsilon) \cap B(x_0, 2\delta^\varepsilon) \tag{4.8}$$

$\forall\ |t - t_0| \le 2\delta^\varepsilon$, and for any x^*, we set

$$D^\varepsilon_n(t : x^*) := \mathcal{R}^2 \setminus \{x^*\} \oplus R^\varepsilon(t) C_n.$$

Then, $D^\varepsilon_n(\cdot : x_0)$ approximates $D^\varepsilon(\cdot : x_0)$, and by (4.4) and (4.8), there are a subsequence n_k and sequences $(x_k, t_k) \to (x_0, t_0)$, $y_k \to x_0^\varepsilon$ satisfying

$$x_k \in \Gamma_{n_k}(t_k) \cap \partial D^\varepsilon_{n_k}(t_k : y_k),$$

$$\Omega_{n_k}(t) \cap B(x_0, \delta^\varepsilon) \subset D^\varepsilon_{n_k}(t : y_k) \cap B(x_0, \delta^\varepsilon) \qquad \forall\ |t - t_0| \le \delta^\varepsilon.$$

Again, we assume that $n_k = k$, write $D_k(t)$ for $D^\varepsilon_{n_k}(t : y_k)$, and let x_k belong to the ith side of $\Gamma_k(t_k)$. Since, in this case, the normal velocity of D_k at t_k is equal to $v - 2K(t_k - t_0)$,

$$V_i = v - 2K(t_k - t_0).$$

If $v \le 0$, (4.3) is immediately satisfied. Hence, we may assume that $v > 0$. So, for small $\varepsilon > 0$, $V_i > 0$, and by (2.3), $\chi_i = -1$. Consequently, $l_i(t_k)$ is greater than or equal to the length of any side of $D_k(t_k)$:

$$l_i(t_k) \ge 2R^\varepsilon(t_k) \sin \frac{\pi}{k},$$

and therefore,

$$v - C(t_k - t_0) = V_i = \frac{2\tan(\pi/k)}{l_i(t_k)} \leq \frac{1}{R^\varepsilon(t_k)\cos(\pi/k)}.$$

We first let $k \to \infty$ and then $\varepsilon \downarrow 0$. Since $R^\varepsilon(t_k)$ converges to $1/|\kappa| = -1/\kappa$, the result is (4.3). □

5. Convergence. Let $\Gamma_0 = \partial\Omega_0$ be a twice differentiable Jordan curve and $\Gamma_{n\,0} = \partial\Omega_{n\,0}$ be an n-smooth approximation of Γ_0 satisfying

$$\lim_{n\to\infty} d_H(\Omega_{n\,0}, \Omega_0) = 0, \tag{5.1}$$

where d_H is the Hausdorff distance. For each n, there is a unique n-smooth solution $\{\Gamma_n(t)\}_{t\in[0,T_n)}$ of (2.3) satisfying the initial condition $\Gamma_n(0) = \Gamma_{n\,0}$ by Theorem 3.1. Moreover,

$$T_n = \frac{|\Omega_{n\,0}|}{2n\tan(\pi/n)} \to T_0 := \frac{|\Omega_0|}{2\pi}, \qquad n \to +\infty. \tag{5.2}$$

Let $\widehat{\Omega}$ and $\underline{\Omega}$ be as in section 4 so that, by construction,

$$\operatorname{cl}\underline{\Omega}(t) \subset \widehat{\Omega}(t) \qquad \forall t \in [0, T_0). \tag{5.3}$$

Moreover, $\widehat{\Omega}$ is a weak subsolution of the mean curvature flow, and $\underline{\Omega}$ is a weak supersolution of the mean curvature flow. In general space dimension, there is no comparison between weak sub- and supersolutions; however, in dimension two, there is always a smooth solution of the mean curvature flow, $\Gamma(t) = \partial\Omega(t)$ and we will show that

$$\widehat{\Omega}(t) \subset \operatorname{cl}\Omega(t) \subset \operatorname{cl}\underline{\Omega}(t) \qquad \forall t \in [0, T_0). \tag{5.4}$$

Combining (5.3) and (5.4), we will then obtain the convergence of Ω_n to Ω in Hausdorff topology, thus generalizing the previous convergence results of Girao [20] and Girao and Kohn [21].

The foregoing outline of our convergence result is entirely analogous to the Barles and Perthame procedure of proving convergence with very weak L^∞ estimates [5, 6].

THEOREM 5.1. *Let $\Gamma_n(t) = \partial\Omega_n(t)$ be the solution of* (2.3) *with initial data $\Gamma_{n\,0}$, and let $\Gamma(t) = \partial\Omega(t)$ be the smooth solution of the mean curvature flow with initial data Ω_0. Assume* (5.1)*; then*

$$\lim_{n\to\infty} d_H(\Omega_n(t), \Omega(t)) = 0 \tag{5.5}$$

locally uniformly in $t \in [0, T_0)$.

We begin with the following lemma.

LEMMA 5.2. $\widehat{\Omega}(0) \subset \operatorname{cl}\Omega_0 \subset \operatorname{cl}\underline{\Omega}(0)$.

Proof. We will prove only the first inclusion. Proof of the second inclusion is similar.

Since $d_H(\Omega_n, \Omega_0) \to 0$, for any $x_0 \in \Omega_0$ there are $\delta_0 > 0$ and $n_0 \in \mathcal{N}$ satisfying

$$B(x_0, \delta_0) \subset\subset \Omega_n \qquad \forall n > n_0.$$

Let γ_n be the regular n-polygon enclosing $B(x_0, \delta_0)$. If necessary, by taking n_0 larger, we may assume that $\gamma_n \subset\subset \Omega_n \ \forall\ n > n_0$. Let $\gamma_n(t)$ be the solution of the crystalline flow (2.3) with initial data $\gamma_n(0) = \gamma_n$ and $\omega_n(t)$ be the open set enclosed by $\gamma_n(t)$. Then by the containment principle for crystalline motions (cf. Giga and Gurtin [18]),

$$B\left(x_0, \delta_0/2\right) \subset \omega_n(t) \subset \Omega_n(t) \qquad \forall n > n_0,\ 0 \le t \le \frac{1}{4}\delta_0^2.$$

Let $n \to +\infty$ and $t \downarrow 0$ to conclude that $B(x_0, \delta_0/2) \subset \underline{\Omega}(0)$; therefore $x_0 \in \underline{\Omega}(0)$. □

In our second step, we will show that the smooth mean curvature flow yields a viscosity sub- and supersolution of the following equation:

$$u_t + F(Du, D^2u) = 0, \qquad \mathcal{R}^2 \times (0, T),$$

where

$$F(p, X) = -\mathrm{tr}((I - \bar{p} \otimes \bar{p})X) \tag{5.6}$$

and $\bar{p} = p/|p|$. This step is very similar to Evans and Spruck [12, Section 6] and Ambrosio and Soner [3, section 3].

We refer to Crandall, Ishii, and Lions [10] and Fleming and Soner [13] for information on viscosity solutions and to Chen, Giga, and Goto [9], and Evans and Spruck [12] for the properties of the level set equations.

Let $\{\Gamma(t)\}_{0 \le t < T_0}$ be a unique smooth mean curvature flow satisfying $\Gamma(0) = \Omega_0$, and let $d(x, t)$ be the signed distance function to $\Gamma(t)$, i.e.,

$$d(x,t) = \begin{cases} \mathrm{dist}(x, \Gamma(t)) & \text{if } x \in \Omega(t), \\ -\mathrm{dist}(x, \Gamma(t)) & \text{otherwise,} \end{cases}$$

where $\Omega(t)$ is the open set enclosed by $\Gamma(t)$. For a scalar d, $d \wedge 0 = \min\{d, 0\}$ and $d \vee 0 = \max\{d, 0\}$.

LEMMA 5.3. *For any $\delta > 0$, there are constants $\sigma = \sigma(\delta) > 0$ and $K = K(\delta) > 0$ so that the function $u(x,t) := e^{-Kt}[(d \vee 0)(x,t) \wedge \sigma]$ is a viscosity subsolution of*

$$u_t + F(Du, D^2u) = 0 \qquad \text{in} \quad \mathcal{R}^2 \times (0, T_0).$$

Proof. For $\delta > 0$, there exists a $\sigma = \sigma(\delta) > 0$ such that d is smooth in $\{x \in \mathcal{R}^2 : |d(x,t)| < 2\sigma\} \times [0, T_0 - \delta]$, and in this tubular set,

$$\Delta d(x,t) = \frac{\kappa(y,t)}{1 - \kappa(y,t)d(x,t)}, \tag{5.7}$$

where $y \in \Gamma(t)$ is a unique point satisfying $|d(x,t)| = |x - y|$ and $\kappa(y,t)$ is the curvature of $\Gamma(t)$ at y. Since $\{\Gamma(t)\}_{0 \le t < T_0}$ is a smooth mean curvature flow,

$$d_t - \Delta d = 0 \qquad \text{in} \quad \Gamma(t) \times (0, T_0). \tag{5.8}$$

Since

$$C(\delta) := \sup\{|\kappa(x,t)| \ :\ (x,t) \in \partial\Omega(t) \times [0, T_0 - \delta]\} < \infty,$$

by (5.7) and (5.8), d is a classical subsolution of

$$d_t - \Delta d - Kd \le 0 \qquad \text{on} \quad \{x \ :\ 0 \le d(x,t) \le 2\sigma\} \times (0, T_0 - \delta]$$

for some $K \geq C(\delta)$. Since $|Dd| = 1$, d is also a classical subsolution of

$$d_t + F(Dd, D^2 d) - Kd = 0 \qquad \text{on} \quad \{x \,:\, 0 \leq d(x,t) \leq 2\sigma\} \times (0, T_0 - \delta].$$

Let h^ϵ be a bounded smooth function satisfying $h^\epsilon(r) = 0$ for $r \leq 0$, $h^\epsilon(r) = \sigma$ for $r \geq \sigma$, and as $\epsilon \downarrow 0$, $h^\epsilon(r)$ converges to $(r \vee 0) \wedge \sigma$. Since F is geometric, i.e.,

$$F(\lambda p, \lambda A + \mu p \otimes p) = \lambda F(p, A), \qquad \lambda, \mu \geq 0,$$

by calculus, we conclude that $u^\epsilon = e^{-Kt} h^\epsilon(d)$ is a classical subsolution of

$$u^\epsilon_t + F(Du^\epsilon, D^2 u^\epsilon) \leq 0 \qquad \text{on} \quad \mathcal{R}^2 \times (0, T_0 - \delta].$$

We let $\epsilon \downarrow 0$, $\delta \downarrow 0$ and use the celebrated stability property of viscosity solutions. □

An entirely similar argument yields the following lemma.

LEMMA 5.4. *For any $\delta > 0$, there are constants $\sigma = \sigma(\delta) > 0$ and $K = K(\delta) > 0$ so that the function $u(x,t) := e^{Kt}[(d \wedge 0)(x,t) \vee (-\sigma)]$ is a viscosity supersolution of*

$$u_t + F(Du, D^2 u) = 0 \qquad \text{in} \quad \mathcal{R}^2 \times (0, T_0).$$

We are now in a position to complete the proof of Theorem 5.1.

Proof of Theorem 5.1. For notational convenience, we set $\Omega_n(t) = \emptyset \;\forall\, n > 1,\ t > T_n$. Let $\widehat{\Omega}$ and $\underline{\Omega}$ be as in section 4, and let $\widehat{T}$, $\underline{T}$ be, respectively, the extinction time of $\widehat{\Omega}(t)$ and $\underline{\Omega}(t)$. Set $\tilde{T} = \min\{\underline{T}, T_0, \widehat{T}\}$.

By Lemma 5.3, $u(x,t) = e^{-Kt}[(d \ \vee 0)(x,t) \wedge \sigma]$ is a viscosity subsolution of

$$u_t + F(Du, D^2 u) = 0 \qquad \text{in} \quad \mathcal{R}^2 \times (0, \tilde{T} - \delta), \tag{5.9}$$

and by Lemma 4.2 and Proposition 6.1, $v(x,t) = \operatorname{dist}(x, \mathcal{R}^2 \backslash \underline{\Omega}(t))$ is a viscosity supersolution of (5.9). Moreover, by Lemma 5.2, $u(\cdot, 0) \leq v(\cdot, 0)$ in $\mathcal{R}^2$, and therefore the comparison principle for solutions of (5.9) (cf. Chen, Giga, and Goto [9], Evans and Spruck [12]) yields

$$u \leq v \qquad \text{in} \quad \mathcal{R}^2 \times [0, \tilde{T} - \delta).$$

We claim that this inequality implies that

$$\Omega(t) \subset \underline{\Omega}(t) \qquad \forall t \in [0, \tilde{T} - \delta).$$

Indeed, for $(x,t) \in \Omega(t) \times [0, \tilde{T} - \delta)$, $0 < u(x,t)$. Then, by the previous inequality, $0 < v(x,t)$ and, therefore, $x \in \underline{\Omega}(t)$.

Similarly, we show that $\widehat{\Omega}(t) \subset \operatorname{cl} \Omega(t) \;\forall\; t \in [0, \tilde{T} - \delta)$, and then we let $\delta \to 0$ to obtain (5.4) on $[0, \tilde{T})$.

A lengthy elementary argument shows that (5.4) is equivalent to (5.5). Hence, (5.5) holds on $[0, \tilde{T})$.

By (5.2) and the construction, $\underline{T} \leq \widehat{T} \leq T_0$. The uniform convergence of Ω_n to Ω implies that $\tilde{T} = T_0$. □

6. Appendix. In this section we gather several properties of the weak solutions.

Let $\{\Omega_n(t)\}_{0 \leq t < T_n}$, $\{\widehat{\Omega}(t)\}_{0 \leq t < T}$, and $\{\underline{\Omega}(t)\}_{0 \leq t < T}$ be as in section 4, and let $d_n(x,t)$ (resp., $\widehat{d}(x,t)$ and $\underline{d}(x,t)$) be the signed distance function for $\{\Omega_n(t)\}_{0 \leq t < T_n}$

(resp., for $\{\widehat{\Omega}(t)\}_{0\le t<T}$ and $\{\underline{\Omega}(t)\}_{0\le t<T}$). Then the definitions of $\widehat{\Omega}(t)$ and $\underline{\Omega}(t)$ are equivalent to

$$(\widehat{d}\wedge 0)(x,t)=\limsup_{\substack{(y,s)\to(x,t)\\ n\to+\infty}}(d_n\wedge 0)(y,s),$$
$$(\underline{d}\vee 0)(x,t)=\liminf_{\substack{(y,s)\to(x,t)\\ n\to+\infty}}(d_n\vee 0)(y,s).$$

The following weak regularity result in t follows from an attendant modification of [29, Lemma 7.3]:

$$\limsup_{y\to x,\ s\uparrow t}(\widehat{d}\wedge 0)(y,s)=(\widehat{d}\wedge 0)(x,t),\qquad (x,t)\in\mathcal{R}^2\times(0,T), \tag{6.1}$$

$$\liminf_{y\to x,\ s\uparrow t}(\underline{d}\vee 0)(y,s)=(\underline{d}\vee 0)(x,t),\qquad (x,t)\in\mathcal{R}^2\times(0,T). \tag{6.2}$$

These identities and the techniques of [29, section 14] yield the equivalence between the weak solutions defined in section 4 and the distance solutions defined by Soner in [29]. Let F be as in (5.6).

PROPOSITION 6.1. *$\{\Omega(t)\}_{0\le t<T}$ is a weak subsolution of the mean curvature flow satisfying* (6.1) *if and only if $d_{\Omega^*}(x,t)\wedge 0$ is a viscosity subsolution of*

$$u_t+F(Du,D^2u)=0\qquad in\quad \mathcal{R}^2\times(0,T). \tag{6.3}$$

$\{\Omega(t)\}_{0\le t<T}$ is a weak supersolution of the mean curvature flow satisfying (6.2) *if and only if $d_{\Omega_*}(x,t)\vee 0$ is a viscosity supersolution of* (6.3).

We close the appendix by proving an approximation result used in section 4.

LEMMA 6.2. *Let $\{O(t)\}_{0\le t<T}$ be a family of closed smooth sets, and let $t_0\in(0,T)$, $x_0\in\partial O(t_0)$ satisfy* (4.4). *Let $D^\varepsilon(t)$ and $D^\varepsilon_n(t:x^*)$ be the same sets as in the proof of* Lemma 4.1. *Assume that $D^\varepsilon(t:x_0^\varepsilon)$ satisfies* (4.7). *Then there are a subsequence n_k and sequences $(x_k,t_k)\to(x_0,t_0)$, $y_k\to x_0^\varepsilon$ as $k\to+\infty$ satisfying*

$$x_k\in\Gamma_{n_k}(t_k)\cap\partial D^\varepsilon_{n_k}(t_k:y_k),$$

$$\Omega_{n_k}(t)\cap B(x_0,\delta^\varepsilon)\subset D^\varepsilon_{n_k}(t:y_k)\cap B(x_0,\delta^\varepsilon)\qquad\forall\ |t-t_0|\le\delta^\varepsilon.$$

Proof. Fix $\varepsilon>0$ and recall $(\widehat{\Omega})^*=\widehat{\Omega}$. Let $d_n(x,t)$ be the signed distance to $D^\varepsilon_n(t:x_0^\varepsilon)$, $d(x,t)$ be the signed distance to $D^\varepsilon(t:x_0^\varepsilon)$, and let

$$\alpha_n:=\inf_{|t-t_0|\le\delta^\varepsilon}\inf\{d_n(x,t)\ :\ x\in\Omega_n(t)\cap B(x_0,\delta^\varepsilon)\}.$$

Choose $t_n\in[t_0-\delta^\varepsilon,t_0+\delta^\varepsilon]$, $x_n\in\Omega_n(t_n)\cap B(x_0,\delta^\varepsilon)$ and $w_n\in\partial D^\varepsilon_n(t_n:x_0^\varepsilon)$ such that

$$|w_n-x_n|=|\alpha_n|.$$

Set

$$y_n=x_0^\varepsilon-(w_n-x_n),$$

so that

$$\Omega_n(t) \cap B(x_0, \delta^\varepsilon) \subset D_n^\varepsilon(t : y_n) \cap B(x_0, \delta^\varepsilon) \quad \forall\ |t - t_0| \le \delta^\varepsilon.$$

Since $x_0 \in \widehat{\Omega}(t_0)$, by the definition of $\widehat{\Omega}$, there are a subsequence n_k and sequences $(z_k, s_k) \to (x_0, t_0)$ such that

$$z_k \in \Omega_{n_k}(s_k).$$

Then

$$\limsup_{k\to\infty} \alpha_{n_k} \le \limsup_{k\to\infty} d_{n_k}(z_k, s_k) = d(x_0, t_0) = 0.$$

A similar argument, using (4.7), shows that $\liminf \alpha_{n_k} \ge 0$. Hence $\alpha_{n_k} \to 0$ and, therefore, $y_{n_k} \to x_0^\varepsilon$.

It remains to show that $(x_{n_k}, t_{n_k}) \to (x_0, t_0)$. Suppose that on a further subsequence, denoted by n_k again,

$$(x_{n_k}, t_{n_k}) \to (\bar{x}, \bar{t}) \in B(x_0, 2\delta^\varepsilon) \times [t_0 - \delta^\varepsilon, t_0 + \delta^\varepsilon].$$

Since d_n converges to d uniformly,

$$d(\bar{x}, \bar{t}) = \lim_{k\to\infty} \alpha_{n_k} = 0 \le \lim_{k\to\infty} d_{n_k}(z_k, s_k) = d(x_0, t_0).$$

Since (x_0, t_0) is the strict minimizer of d, this implies that $(\bar{x}, \bar{t}) = (x_0, t_0)$. □

REFERENCES

[1] F. Almgren and J. E. Taylor, *Flat flow is motion by crystalline curvature for curves with crystalline energies*, J. Differential Geom., 42 (1995), pp. 1–22.

[2] F. Almgren, J. E. Taylor, and L. Wang, *Curvature-driven flows: A variational approach*, SIAM J. Control Optim., 31 (1993), pp. 387–438.

[3] L. Ambrosio and H. M. Soner, *Level set approach to mean curvature flow in arbitrary codimension*, J. Differential Geom., 43 (1996), pp. 693–737.

[4] S. Angenent and M. E. Gurtin, *Multiphase thermomechanics with interfacial structure 2. Evolution of an isothermal interface*, Arch. Rat. Mech. Anal., 108 (1989), pp. 323–391.

[5] G. Barles and B. Perthame, *Discontinuous solutions of deterministic optimal stopping problems*, Math. Model. Numer. Anal., 21 (1987), pp. 557–579.

[6] G. Barles and B. Perthame, *Exit time problems in optimal control and vanishing viscosity method*, SIAM J. Control Optim., 26 (1988), pp. 1133–1148.

[7] G. Barles, H. M. Soner, and P. E. Sougandis, *Front propagation and phase field theory*, SIAM J. Control Optim., 31 (1993), pp. 439–469.

[8] K. A. Brakke, *The Motion of a Surface by Its Mean Curvature*, Princeton University Press, Princeton, NJ, 1978.

[9] Y.-G. Chen, Y. Giga, and S. Goto, *Uniqueness and existence of viscosity solutions of generalized mean curvature flow equations*, J. Differential Geom., 33 (1991), pp. 749–786.

[10] 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.

[11] M. G. Crandall and P.-L. Lions, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–43.

[12] L. C. Evans and J. Spruck, *Motion of level sets by mean curvature*, J. Differential Geom., 33 (1991), pp. 635–681.

[13] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[14] T. Fukui and Y. Giga, *Motion of a graph by nonsmooth weighted curvature*, in Proc. First World Cong. of Nonlinear Anal. 92, V. Lakshmikantham, ed., Walter de Gruyter, Berlin, 1996, pp. 47–56.

[15] M. Gage and R. Hamilton, *The heat equation shrinking convex plane curves*, J. Differential Geom., 23 (1986), pp. 69–95.
[16] M.-H. Giga and Y. Giga, *Evolving graphs by singular weighted curvature*, Arch. Rational. Mech. Anal., 141 (1998), pp. 117–198.
[17] M.-H. Giga and Y. Giga, *Stability for Evolving Graphs by Nonlocal Weighted Curvature*, preprint, Hokkaida University, Japan, 1996.
[18] Y. Giga and M. E. Gurtin, *A comparison principle for crystalline evolution in the plane*, Quat. Appl. Math., 54 (1996), pp. 727–737.
[19] Y. Giga, M. E. Gurtin, and J. Mathias, *On the dynamics of crystalline motions*, Japan. J. Indust. Appl. Math., 15 (1998), pp. 7–50.
[20] P. M. Girão, *Convergence of a crystalline algorithm for the motion of a simple closed convex curve by weighted curvature*, SIAM J. Numer. Anal., 32 (1995), pp. 886–899.
[21] P. M. Girão and R. V. Kohn, *Convergence of a crystalline algorithm for the heat equation in one dimension and for the motion of a graph by weighted curvature*, Numer. Math., 67 (1994), pp. 41–70.
[22] M. E. Gurtin, *Thermodynamics of Evolving Phase Boundaries in the Plane*, Oxford University Press, Oxford, 1993.
[23] M. E. Gurtin, H. M. Soner, and P. E. Souganidis, *Anisotropic motion of an interface relaxed by the formation of infinitesimal wrinkles*, J. Differential Equations, 119 (1995), pp. 54–108.
[24] M. A. Grayson, *The heat equation shrinks embedded plane curves to round points*, J. Differential Geom., 26 (1987), pp. 285–314.
[25] G. Huisken, *Flow by mean curvature of convex surfaces into spheres*, J. Differential Geom., 20 (1984), pp. 237–266.
[26] S. Osher and J. Sethian, *Front propagating with curvature depending speed*, J. Comput. Phys., 79 (1988), pp. 12–49.
[27] M. Ohnuma and M.-H. Sato, *Singular degenerate parabolic equations with applications to geometric evolutions*, Differential Integral Equations, 6 (1993), pp. 1265–1280.
[28] T. Ohta, D. Jasnow, and K. Kawasaki, *Universal scaling in the motion of a random interface*, Phys. Rev. Lett., 49 (1982), pp. 1223–1226.
[29] H. M. Soner, *Motion of a set by the curvature of its boundary*, J. Differential Equations, 101 (1993), pp. 313–372.
[30] H. M. Soner, *Front Propagation*, CRM Proc. Lecture Notes 13, AMS, Providence, RI, 1998, pp. 185–206.
[31] J. E. Taylor, *On the global structure of crystalline surfaces*, Discrete. Comput. Geometry, 6 (1991), pp. 225–262.
[32] J. E. Taylor, *Mean curvature and weighted mean curvature*, Acta. Metal., 40 (1992), pp. 1475–1485.
[33] J. E. Taylor, *Motion of curves by crystalline curvature, including triple junctions and boundary points*, Proc. Sympos. Pure Math., 54 (1993), pp. 417–438.
[34] J. E. Taylor, J. W. Cahn, and A. C. Handwerker, *Geometric models of crystal growth*, Acta. Metal., 40 (1992), pp. 1443–1474.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 38–52

ON L^1 CONVERGENCE RATE OF VISCOUS AND NUMERICAL APPROXIMATE SOLUTIONS OF GENUINELY NONLINEAR SCALAR CONSERVATION LAWS*

WEI-CHENG WANG†

Abstract. We study the rate of convergence of the viscous and numerical approximate solution to the entropy solution of genuinely nonlinear scalar conservation laws with piecewise smooth initial data. We show that the $O(\epsilon|\log\epsilon|)$ rate in L^1 is indeed optimal for viscous Burgers equation. Through the Hopf–Cole transformation, we can study the detailed structure of $\|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1}$. For centered rarefaction wave, the $O(\epsilon|\log\epsilon|)$ error occurs on the edges where the inviscid solution has a corner, and persists as long as the edges remain. The $O(\epsilon|\log\epsilon|)$ error must also occur at the critical time when a new shock forms automatically from the decreasing part of the initial data; thus it is, in general, impossible to maintain $O(\epsilon)$ rate for all $t>0$. In contrast to the centered rarefaction wave case, the $O(\epsilon|\log\epsilon|)$ error at critical time is transient. It resumes the $O(\epsilon)$ rate right after the critical time due to nonlinear effect. Similar examples of some monotone schemes, which admit a discrete version of the Hopf–Cole transformation, are also included.

Key words. hyperbolic conservation laws, error estimates, viscosity methods, monotone schemes

AMS subject classifications. 65M10, 65M05, 35L65

PII. S0036141097316408

1. Introduction.

The hyperbolic conservation law

$$\begin{aligned} &u_t + f(u)_x = 0, \quad x\in\mathbf{R}, \quad t>0,\\ &u(x,0)=u_0(x) \end{aligned} \tag{1.1}$$

can be analyzed using the method of characteristics. Due to nonlinearity of f, the characteristic lines can intersect each other in finite time, and the solution develops jump discontinuities even if the initial data is smooth. Due to the presence of jump discontinuities, we need to generalize the solution class to include "weak solutions." In addition, since the weak solutions are not unique, entropy conditions are needed to specify physically meaningful weak solutions.

There are several equivalent forms of the entropy condition for genuinely nonlinear (say, $f''>0$) scalar conservation laws. Among them is the method of vanishing viscosity, which asserts that the physically relevant solution is obtained by solving the following viscous approximate equation

$$\begin{aligned} &u^\epsilon_t + f(u^\epsilon)_x = \epsilon u^\epsilon_{xx}, \quad x\in\mathbf{R}, \quad t>0, \quad \epsilon>0,\\ &u^\epsilon(x,0)=u_0(x) \end{aligned} \tag{1.2}$$

and letting ϵ go to zero. It is known that u^ϵ converges strongly, and the limiting function, u, is a weak solution of (1.1). Furthermore, u is the unique solution that satisfies the following entropy condition:

$$\frac{u(x+a,t)-u(x,t)}{a}\le\frac{E}{t}, \quad t>0, \tag{1.3}$$

*Received by the editors February 7, 1997; accepted for publication (in revised form) October 15, 1997; published electronically September 25, 1998.
http://www.siam.org/journals/sima/30-1/31640.html
†Courant Institute of Mathematical Sciences, New York University, 251 Mercer Street, New York, NY 10012. Current address: MSRI, 1000 Centennial Drive, Berkeley, CA 94720 (wangwc@msri.org).

where E is a constant depending only on the flux function f and initial data (see, for example, [11] for details).

Monotone difference schemes are first-order numerical schemes used to compute approximate solutions of (1.1):

$$w_j^{n+1} = G(w_{j-p}^n, w_{j-p+1}^n, \dots, w_{j+q}^n) = w_j^n - \lambda[\bar{f}(w_{j-p+1}^n, \dots, w_{j+q}^n) - \bar{f}(w_{j-p}^n, \dots, w_{j+q-1}^n)], \tag{1.4}$$

where p and q are fixed nonnegative integers, G is a monotonely nondecreasing function in each of its arguments, $\bar{f}$ is a Lipchitz continuous function and is consistent with the scalar conservation law (1.1) in the sense

$$\bar{f}(w, \dots, w) = f(w), \tag{1.5}$$

and $\lambda = \Delta t/\Delta x$ is a constant satisfying the CFL condition $\lambda < |f'|$.

Most well-known first-order schemes such as the Lax–Friedrichs scheme, Godunov scheme, and Enquist–Osher scheme are monotone schemes. Monotone schemes are known to converge to the entropy solution of (1.1) as $\Delta x \to 0$ (see [2]) and they are at most first-order accurate [1].

Whether or not a viscous approximation/monotone scheme can be of order $O(\epsilon)/O(\Delta x)$ accurate is an issue of practical interest and has long been studied. Although viscous approximation and monotone schemes are formally first order, they can really lose half-order accuracy across discontinuities. For example, it is easy to see, using a scaling argument, that the solution of the heat equation with an initial jump discontinuity is indeed $O(\sqrt{\epsilon})$ in L^1 norm away from its zero viscosity limit. In fact, Tang and Teng [14] proved that the $O(\sqrt{\epsilon})$ or $O(\sqrt{\Delta x})$ rate is indeed optimal for *all* monotone schemes applied to linear advection equations with discontinuous data.

For general BV initial data with genuinely nonlinear flux, several authors have obtained $O(\sqrt{\epsilon})$ or $O(\sqrt{\Delta x})$ rate. See, for example, Kuznetsov [6], Lucier [8], Sanders [10], and Tadmor [13]. It turns out to be optimal for this case (i.e., beyond linear degeneracy); see Sabac [9]. For the special case of monotonely nondecreasing initial data, Harabetian [5] has obtained $O(\epsilon|\log\epsilon|)/O(\Delta x|\log(\Delta x)|)$ rate in L^1 norm and showed that it is indeed optimal in this case.

Although BV solution is a natural class for genuinely nonlinear scalar conservation law, we will consider here only the subclass of piecewise smooth solutions with finitely many shocks. This class is of practical interest in shock capturing for the following reason: We expect viscous solution or monotone schemes to have better resolutions across an isolated jump discontinuity if the flux function is genuinely nonlinear, since for a linearly degenerate flux function, the discontinuity is a contact one, thus the smearing is a result of diffusion only; while in case of a genuinely nonlinear flux, the entropy condition dictates that the characteristic curves impinge into the shock, and thus tend to squeeze the profile in shape.

The first result in this direction was Goodman and Xin [4], where the authors considered piecewise smooth flows with noninteracting shocks for *systems* of hyperbolic conservation laws with viscosity. They obtained an $O(\epsilon)$ estimate away from shock regions and an overall $O(\epsilon^\gamma)$ rate for any $\gamma < 1$. The proof uses a matched asymptotic analysis employing a superposition of outer solutions (asymptotic series off the shock) and inner solutions (asymptotic expansion near the shock in stretched variables), as well as a nonlinear stability analysis based on energy estimates.

Inspired by [4], Teng and Zhang [16], Tang and Teng [15] showed that, for genuinely nonlinear scalar conservation laws with piecewise smooth initial data having

finitely many inflection points, the convergence rate can be improved to $\sup_{t>0} \|u(\cdot,t) - u^\epsilon(\cdot,t)\| \le O(\epsilon|\log\epsilon|)$. And in case there is no centered rarefaction wave or shock formation at a later time, the rate is actually $O(\epsilon)$.

The corresponding counterpart for monotone schemes is more subtle. Engquist and Yu [3] and Smyrlis and Yu [12] obtained pointwise estimates for a wide class of finite difference schemes, which result in the $O(\Delta x)$ convergence rate in L^1 norm of monotone schemes for piecewise smooth initial data with noninteracting shocks provided no shocks form at a later time.

From the argument in [15], it is not clear whether the $O(\epsilon|\log\epsilon|)$ is optimal beyond the centered rarefaction wave case (the optimality of the centered rarefaction wave case was shown in Harabetian [5]). In this paper, we will study in detail the structure of $\|u(\cdot,t) - u^\epsilon(\cdot,t)\|$ through an example, the viscous Burgers equation, to gain more insight. It turns out that this rate is actually obtained at the critical time when a shock develops from the decreasing part of the initial data. Thus, it is in general impossible to maintain $O(\epsilon)/O(\Delta x)$ rate for all $t > 0$. However, in contrast to the centered rarefaction wave case, this phenomenon is transient; it resumes the $O(\epsilon)/O(\Delta x)$ rate right after the critical time. (This case was not covered in [14], [15], [3], and [12], where the authors considered the shocks coming from jumps in initial data.) This result is consistent with the following heuristic argument: The viscous approximation/monotone schemes are first-order accurate both before and after the critical time, but for different reasons. Before the critical time, the solution of (1.1) is smooth if the initial data is; therefore, the viscous term ϵu^ϵ_{xx} is an $O(\epsilon)\cdot O(1)$ quantity. After the critical time, the shock is already formed, and the impinging of characteristic lines counteract with diffusion. However, at the critical time, neither of these mechanisms is available, resulting in an underresolution.

The rest of this paper is arranged as follows: In section 2, we will review some basic facts about formation of shocks, the Hopf–Cole transformation, and a few lemmas to be used later. In section 3.1, we state and prove the main theorem concerning the convergence rate at and after critical time for the viscous Burgers equation using the Hopf–Cole transformation. In section 3.2, we give the same results for several monotone schemes with particular flux functions, including upwind, Lax–Friedrichs, and Godunov scheme, which admit a discrete version of Hopf–Cole transformation. It will be clear how these elementary arguments can be utilized to study the centered rarefaction wave case, and interactions of shocks and centered rarefaction waves, etc. The results are as stated in the abstract of this paper; we thus omit the details.

2. Preliminaries.

Notation: $\|\cdot\|$ is the L^1 norm. We'll also denote the local L^1 integral $\int_a^b |g(x)|dx$ by $\|g\|_{L^1(dx;[a,b])}$.

Consider the viscous and inviscid Burgers equation which are special cases of (1.1) and (1.2) with $f(u) = \frac{1}{2}u^2$,

$$\begin{aligned} &u_t + uu_x = 0, \quad x \in \mathbf{R}, \quad t > 0, \\ &u(x,0) = u_0(x) \end{aligned} \tag{2.1}$$

and

$$\begin{aligned} &u^\epsilon_t + u^\epsilon u^\epsilon_x = \epsilon u^\epsilon_{xx}, \quad x \in \mathbf{R}, \quad t > 0, \quad \epsilon > 0, \\ &u^\epsilon(x,0) = u_0(x). \end{aligned} \tag{2.2}$$

We first recall some facts about spontaneous formation of shocks. If the initial data is smooth and is such that $f'(u_0(\cdot))$ is not monotonely nondecreasing, the

characteristic lines can intersect each other and the shock forms. If $\xi_1 < \xi_2$ with $f'(u_0(\xi_1)) > f'(u_0(\xi_2))$, then the two characteristic lines starting from ξ_1 and ξ_2 intersect at time $t = \frac{\xi_2-\xi_1}{f'(u_0(\xi_1))-f'(u_0(\xi_2))}$; thus the first time at which neighboring characteristic lines intersect is when $t = t_c = \frac{-1}{\min_\xi \frac{d}{d\xi} f'(u_0(\xi))}$ and the initial shock is located at the characteristic line starting from ξ_0 where the minimum is taken.

Here in Burgers's equation, $f'(u_0) = u_0$. Up to a Galilean transformation, we may assume that $u_0(0) = 0$ and that $\xi = 0$ is where u_0' assumes its negative minimum which corresponds to the initial formation of the shock. Thus the local Taylor expansion near $\xi = 0$ reads

$$u_0(\xi) = -\frac{1}{t_c}\xi + a\xi^{2p+1} + \cdots, \tag{2.3}$$

where $a > 0$ and p is a positive integer. We'll carry out the analysis for $p = 1$; the proof for other values of p is similar.

By differentiating (2.1) with respect to x and then integrating along the characteristic lines, one can find that the derivative blows up near $t = t_c$ (for a detail derivation, see, for example, [11]),

$$u_x(0,t) = -\frac{O(1)}{t_c - t},\ t < t_c. \tag{2.4}$$

Our main tool is the classical Hopf–Cole transformation,

$$u^\epsilon = -2\epsilon(\log \phi^\epsilon)_x. \tag{2.5}$$

Through (2.5), (2.2) linearizes to the heat equation

$$\phi_t^\epsilon = \epsilon\phi_{xx}^\epsilon. \tag{2.6}$$

After transforming the initial data and solving the heat equation, we have

$$u^\epsilon(x,t) = -2\epsilon\left(\log \int_{-\infty}^{\infty} e^{-\frac{1}{2\epsilon}G(x,y,t)}dy\right)_x, \tag{2.7}$$

where $G(x,y,t) = \int_0^y u_0(y')dy' + \frac{(x-y)^2}{2t}$.

Since (2.7) gives an exact formula for $u^\epsilon(x,t)$, hence for $\int^x u^\epsilon(x',t)dx'$, we can estimate $\|u^\epsilon(\cdot,t) - u(\cdot,t)\|$ as long as we know the sign of $u^\epsilon(\cdot,t) - u(\cdot,t)$. The following lemma is based on this observation.

LEMMA 2.1. *Let u_0 be a smooth and bounded function satisfying*

(A.1) *$u_0(\xi) = -\frac{\xi}{t_c} + a\xi^3 + b\xi^4 + O(\xi^5)$ for $|\xi| < \delta$ where $a > 0$;*

(A.2) *$\xi = 0$ is the point corresponding to the first spontaneous formation of shocks, that is, $u_0'(\xi) > -\frac{1}{t_c}$ for all $\xi \neq 0$;*

(A.3) *u_0 is antisymmetric: $u_0(-\xi) = -u_0(\xi)$;*

(A.4) *u_0 is monotonely decreasing;*

(A.5) *u_0 is concave on $\xi < 0$, and, therefore, by Assumption* (A.3)*, convex on $\xi > 0$.*

Then

$$u(x,t_c) \geq (\leq) u^\epsilon(x,t_c) \quad \text{on } x < (>) 0. \tag{2.8}$$

Proof. By symmetry, we only need to prove the statement on $\{x < 0\}$. We will apply the maximum principle in the region $\{(x,t) : 0 < t < t_c, x < 0\}$ for $w = u^\epsilon - u$, which satisfies

$$w_t + (aw)_x - \epsilon w_{xx} = \epsilon u_{xx}, \tag{2.9}$$

where $a(x,t) = \frac{1}{2}(u^\epsilon(x,t) + u(x,t)) = \frac{1}{2}w(x,t) + u(x,t)$. Clearly, $w = 0$ on $\{(x,t) : t = 0, x < 0\}$ by definition and on $\{(x,t) : 0 < t < t_c, x = 0\}$ by symmetry. Since monotonicity and concavity of u is preserved under the characteristic flow (one can see this by differentiating (1.1) once and twice, then integrating along the characteristic lines), we have the correct signs on the right-hand side of (2.8) and the coefficient of w in order to apply the maximum principle, by which we conclude that $w \leq 0$ on $\{x < 0\}$. □

We'll also need the following lemma.

LEMMA 2.2 (L^1 stability). *If $u_i^\epsilon(x,t)$, $i = 1, 2$ satisfy*

$$\frac{\partial}{\partial t}u_i^\epsilon + \frac{\partial}{\partial x}f(u_i^\epsilon) - \epsilon\frac{\partial^2}{\partial x^2}u_i^\epsilon = g_i(x,t), \tag{2.10}$$

then

$$\|u_1^\epsilon(\cdot,t) - u_2^\epsilon(\cdot,t)\| \leq \|u_1^\epsilon(\cdot,0) - u_2^\epsilon(\cdot,0)\| + \int_0^t \|g_1(\cdot,s) - g_2(\cdot,s)\|ds. \tag{2.11}$$

Proof. Let $w = u_1^\epsilon - u_2^\epsilon$; then w solves the following equation:

$$w_t + (aw)_x - \epsilon w_{xx} = g_1 - g_2, \tag{2.12}$$

where $a(x,t) = \frac{1}{2}(u_1^\epsilon + u_2^\epsilon)$ for Burgers's equation. (For general flux, $a(x,t)$ is a proper average of $f'(u_1^\epsilon(x,t))$ and $f'(u_2^\epsilon(x,t))$.) Since the backward adjoint equation

$$\begin{aligned} &z_t + az_x + \epsilon z_{xx} = 0, \\ &z(\cdot,t) = \operatorname{sgn}(w(\cdot,t)) \end{aligned} \tag{2.13}$$

satisfies the maximum principle, we then complete the proof by integrating $z \cdot (2.12) + w \cdot (2.13)$ by parts. □

3. Convergence rate at and near critical time.

3.1. The Burgers equation.

THEOREM 3.1. *Let $u^\epsilon(x,t)$ and $u(x,t)$ be solutions of* (2.1) *and* (2.2)*, respectively, with the same initial data $u_0(x)$ satisfying* (A.1) *and* (A.2) *in Lemma* 2.1*, then for t near t_c, we have*

1. *If $t \neq t_c$, then*

$$\|u^\epsilon(\cdot,t) - u(\cdot,t)\| \leq C(t)\epsilon \quad \text{as} \quad \epsilon \to 0,$$

where $C(t) = O(\log \frac{1}{|t-t_c|})$.

2.

$$\|u^\epsilon(\cdot,t_c) - u(\cdot,t_c)\| = O(\epsilon|\log\epsilon|) \quad \text{as} \quad \epsilon \to 0.$$

Proof. The case $t < t_c$ of the first part is a direct consequence of Lemma 2.2 above, since $\|u_{xx}(\cdot,t)\| = TV(u_x(\cdot,t)) = 2\|u_x(\cdot,t)\|_{L^\infty} = -2u_x(0,t) = O(\frac{1}{t_c - t})$.

At $t = t_c$, we first prove the special case where the initial data satisfy the assumptions of Lemma 2.1. In this case we see that from (2.8),

$$\|u^\epsilon(\cdot,t_c) - u(\cdot,t_c)\|_{L^1(dx;[-1,0])} = \int_{-1}^0 u(x,t_c)dx - \int_{-1}^0 u^\epsilon(x,t_c)dx. \tag{3.1.1}$$

By the Hopf–Cole transformation,

$$\int_{-1}^{0} u^\epsilon(x,t)dx = -2\epsilon \log \left(\frac{\int e^{-\frac{1}{2\epsilon}G(0,y,t)}dy}{\int e^{-\frac{1}{2\epsilon}G(-1,y,t)}dy} \right), \tag{3.1.2}$$

where $G(x,y,t) = \int_0^y u_0(y')dy' + \frac{(x-y)^2}{2t}$ and the domain of integration in $int(\cdot)dy$ is the whole real line. Following the standard stationary phase method, we check that $G_{yy}(-1,\xi(-1,t_c),t_c) = u_0'(\xi) + \frac{1}{t_c} > 0$, where $\xi = \xi(x,t)$ is where $G(x,\cdot,t)$ assumes its global minimum,

$$u_0(\xi(x,t)) = \frac{x-\xi(x,t)}{t}. \tag{3.1.3}$$

Thus at $t = t_c$, the leading-order asymptotic expansion of the denominator in (3.1.2) is

$$\begin{aligned}
\int e^{-\frac{1}{2\epsilon}G(-1,y,t_c)}dy &= e^{-\frac{1}{2\epsilon}G(-1,\xi(-1,t_c),t_c)} \int e^{-\frac{1}{2\epsilon}[G(-1,y,t_c)-G(-1,\xi(-1,t_c),t_c)]}dy \\
&\sim e^{-\frac{1}{2\epsilon}G(-1,\xi(-1,t_c),t_c)} \int e^{-\frac{1}{2\epsilon}\frac{G_{yy}(-1,\xi(-1,t_c),t_c)}{2}(y-\xi(-1,t_c))^2}dy \\
&= \frac{2\sqrt{\pi}}{u_0'(\xi(-1,t_c)) + \frac{1}{t_c}} \cdot \epsilon^{\frac{1}{2}} \exp\left(-\frac{1}{2\epsilon}G(-1,\xi(-1,t_c),t_c) \right). \qquad (3.1.4)
\end{aligned}$$

The numerator, however, has a quartic exponent $G(0,y,t_c)$ at $(x,t) = (0,t_c)$, since $\xi(0,t_c) = 0$, $G_y(0,\xi(0,t_c),t_c) = G_{yy}(0,\xi(0,t_c),t_c) = G_{yyy}(0,\xi(0,t_c),t_c) = 0$ and $G_{yyyy}(0,\xi(0,t_c),t_c) = 6a > 0$. Thus the asymptotic expansion of the integral is, to leading order,

$$\begin{aligned}
\int e^{-\frac{1}{2\epsilon}G(0,y,t_c)}dy &= e^{-\frac{1}{2\epsilon}G(0,0,t_c)} \int e^{-\frac{1}{2\epsilon}[G(0,y,t_c)-G(0,0,t_c)]}dy \\
&\sim e^{-\frac{1}{2\epsilon}G(0,0,t_c)} \int e^{-\frac{1}{8\epsilon}ay^4}dy \\
&= I_0 (\frac{4\epsilon}{a})^{\frac{1}{4}} \exp\left(-\frac{1}{2\epsilon}G(0,0,t_c) \right), \qquad (3.1.5)
\end{aligned}$$

where $I_0 = \int_{-\infty}^{\infty} e^{-\frac{z^4}{2}}dz$ is a constant. Therefore,

$$\int_{-1}^{0} u^\epsilon(x,t_c)dx \sim G(0,0,t_c) - G(-1,\xi(-1,t_c),t_c) + \frac{1}{2}\epsilon\log\epsilon + \cdots. \tag{3.1.6}$$

By differentiating (3.1.3) with respect to x, we see that $\frac{\partial}{\partial x}G(x,\xi(x,t),t) = u(x,t)$, so

$$G(0,0,t_c) - G(-1,\xi(-1,t_c),t_c) = \int_{-1}^{0} u(x,t_c)dx. \tag{3.1.7}$$

From (3.1.2), (3.1.4), (3.1.5), (3.1.6), and (3.1.7), we conclude that

$$\|u^\epsilon(\cdot,t_c) - u(\cdot,t_c)\|_{L^1(dx;[-1,0])} \sim \frac{1}{2}\epsilon|\log\epsilon|. \tag{3.1.8}$$

The same estimate holds for $\|u^\epsilon(\cdot,t_c)-u(\cdot,t_c)\|_{L^1(dx;[0,1])}$ by symmetry. The integral outside of $[-1,1]$ is of lower order by virtue of Lemma 3.2 below. Thus the special case is proved.

To prove the general case, we note that, because of the structure of the initial data, the assumptions of Lemma 2.1, except (A.3), indeed hold for ξ near zero in general. Thus we only have to take care of the antisymmetry. We proceed as follows.

Let $\delta_0>0$ be a small number such that all the assumptions in Lemma 2.1, except (A.3), are valid for $|\xi|<2\delta_0$, and let the characteristic line starting from $(-\delta_0,0)$ intersect the line $\{t=t_c\}$ at $(-\delta_1,t_c)$. We will concentrate on the local deviation $\|u^\epsilon(\cdot,t_c)-u(\cdot,t_c)\|_{L^1(dx;[-\delta_1,0])}$. The following lemma allows us to modify the initial data in order to reduce to the antisymmetric case.

LEMMA 3.2. *Let v_0, w_0 be two bounded initial data such that $v_0(\xi)=w_0(\xi)$ on $\{\xi\ge\alpha\}$ for some $\alpha\in\mathbf{R}$ and let $v^\epsilon(x,t)$, $w^\epsilon(x,t)$ be corresponding solutions of the viscous Burgers equation. If for some $\beta>\alpha$, the characteristic flows of v_0 and w_0 left of α are strictly separated from those right of β up to some time $t_1>0$, that is, if there exist $\beta_1>\alpha_1$, such that the characteristic lines of v_0 and w_0 starting from left of $(\alpha,0)$ intersect the line $\{t=t_1\}$ at left of (α_1,t_1), and vice versa on the right of $(\beta,0)$ and (β_1,t_1), then*

$$\|v^\epsilon(\cdot,t_1)-w^\epsilon(\cdot,t_1)\|_{L^1(dx;[\beta_1,\infty))}=O(1)e^{-\frac{O(1)}{\epsilon}}. \tag{3.1.9}$$

Proof. Equation (3.1.9) can be proved by estimating the Green function of the backward adjoint equation, or one can prove it directly using the Hopf–Cole transformation. □

Therefore, by adjusting $u_0(\xi)$ on $\{\xi<-2\delta_0\}$ if necessary, we may assume that u_0 satisfies the assumptions (A.2), (A.4), and (A.5) globally on $\{\xi<0\}$. To reduce to the special case, we now construct an antisymmetric initial data

$$u_{a,0}(\xi)=\begin{cases}u_0(\xi) & \text{if } \xi<0,\\ -u_0(-\xi) & \text{if } \xi\ge 0.\end{cases} \tag{3.1.10}$$

Since the corresponding inviscid solutions agree on the interval under consideration,

$$u_a(x,t_c)=u(x,t_c)\quad\text{on}\quad -\delta_1\le x\le 0, \tag{3.1.11}$$

it suffices to estimate $\|u_a^\epsilon(\cdot,t_c)-u^\epsilon(\cdot,t_c)\|_{L^1(dx;[-\delta_1,0])}$.

By Assumption (A.1), we have $u_{a,0}(\xi)-u_0(\xi)\le(\ge)0$ on $\{\xi\le 2\delta_0\}$ if $b>(<)0$. By Lemma 3.2, we can adjust $u_0(\xi)$ on $\{\xi>2\delta_0\}$ if necessary, so we may assume, without loss of generality, that

$$u_{a,0}(\xi)-u_0(\xi)\le(\ge)0\ \text{ for all }\ \xi\in\mathbf{R}\ \text{ if }\ b>(<)0. \tag{3.1.12}$$

From the classical comparison lemma [2], (3.1.12) implies that

$$u_a^\epsilon(x,t)-u^\epsilon(x,t)\le(\ge)0\ \text{ if }\ b>(<)0, \tag{3.1.13}$$

and, therefore,

$$\|u_a^\epsilon(\cdot,t_c)-u^\epsilon(\cdot,t_c)\|_{L^1(dx;[-\delta_1,0])}=\left|\int_{-\delta_1}^0 u_a^\epsilon(x,t_c)dx-\int_{-\delta_1}^0 u^\epsilon(x,t_c)dx\right| \tag{3.1.14}$$

Equations (3.1.11), (3.1.13), and (3.1.14) together imply

$$\begin{aligned}&\|u^\epsilon(\cdot,t_c)-u(\cdot,t_c)\|_{L^1(dx;[-\delta_1,0])},\\ &=\begin{cases}\displaystyle\int_{-\delta_1}^0 u(x,t_c)dx-\int_{-\delta_1}^0 u^\epsilon(x,t_c)dx & \text{if } b<0,\\ \displaystyle\int_{-\delta_1}^0 u(x,t_c)dx+\int_{-\delta_1}^0 u^\epsilon(x,t_c)dx-2\int_{-\delta_1}^0 u_a^\epsilon(x,t_c)dx & \text{if } b>0;\end{cases}\end{aligned}\tag{3.1.15}$$

therefore, we can apply the Hopf–Cole transformation again,

$$\begin{aligned}&\|u^\epsilon(\cdot,t_c)-u(\cdot,t_c)\|_{L^1(dx;[-\delta_1,0])},\\ &=\begin{cases}\displaystyle -2\epsilon\log\left(\frac{\int e^{-\frac{1}{2\epsilon}G(-\delta_1,y,t_c)-G(-\delta_1,\xi(-\delta_1,t_c),t_c)}dy}{\int e^{-\frac{1}{2\epsilon}G(0,y,t_c)-G(0,0,t_c)}dy}\right) & \text{if } b<0,\\ \displaystyle -2\epsilon\Bigg[\log\left(\frac{\int e^{-\frac{1}{2\epsilon}G_a(-\delta_1,y,t_c)-G_a(-\delta_1,\xi(-\delta_1,t_c),t_c)}dy}{\int e^{-\frac{1}{2\epsilon}G_a(0,y,t_c)-G_a(0,0,t_c)}dy}\right)\\ \displaystyle +\log\left(\frac{\int e^{-\frac{1}{2\epsilon}G(0,y,t_c)}dy}{\int e^{-\frac{1}{2\epsilon}G_a(0,y,t_c)}dy}\right)\Bigg] & \text{if } b>0,\end{cases}\end{aligned}\tag{3.1.16}$$

where $G_a(x,y,t)=\int_0^y u_{a,0}(y')dy'+\frac{(x-y)^2}{2t}$, and $G_a(-\delta_1,\xi(-\delta_1,t_c),t_c)=G(-\delta_1,\xi(-\delta_1,t_c),t_c)$ cancel out in the second term of the case $b>0$ in (3.1.17). Since $u_0'''(0)$ is preserved under antisymmetrization, we see that from (3.1.5)

$$\frac{\int e^{-\frac{G(0,y,t_c)}{2\epsilon}}dy}{\int e^{-\frac{G_a(0,y,t_c)}{2\epsilon}}dy}=1+o(1).\tag{3.1.17}$$

In view of (3.1.17) and (3.1.17), we have

$$\|u^\epsilon(\cdot,t_c)-u(\cdot,t_c)\|_{L^1(dx;[-\delta_1,0])}\sim\frac{1}{2}\epsilon|\log\epsilon|.\tag{3.1.18}$$

The estimate for $\|u^\epsilon(\cdot,t_c)-u(\cdot,t_c)\|_{L^1(dx;[0,\delta_1])}$ is similar. The general case for $t=t_c$ is thus proved.

The case $t>t_c$ can be reduced to the case $t<t_c$ by constructing a new initial data which delays the formation of the shock. Let $t_0>t_c$ be given, with t_0-t_c sufficiently small. Denote by $s(t)$ the location of the shock at time t, and let $(\xi_-,0)$ be where the backward characteristic line from $(s(t_0)-0,t_0)$ intersects the x-axis. For t_0-t_c sufficiently small, ξ_- is close to 0 and the tangent line of $u_0(\cdot)$ at $(\xi_-,u_0(\xi_-))$ lies above u_0 in a neighborhood of ξ_-. Now define

$$\bar u_0(\xi)=\begin{cases}u_0(\xi) & \text{if } \xi<\xi_-,\\ \max(u_0(\xi),u_0(\xi)+u_0'(\xi_-)(\xi-\xi_-)\ ) & \text{if } \xi\ge\xi_-,\end{cases}$$

and let $\bar u(x,t)$ and $\bar u^\epsilon(x,t)$ be correponding inviscid and viscous solutions. It is easy to see that

(a) $\bar{u}(x,t_0) = u(x,t_0)$ for $x < s(t_0)$.
(b) The critical time for $\bar{u}_0$ is $\bar{t}_c = -\frac{1}{u_0'(\xi_-)} > t_0$; thus no shock forms in $\bar{u}(\cdot,\cdot)$ up to $t = t_0$. Moreover, $\bar{t}_c - t_0 = \frac{1}{2}(t_0 - t_c) + O((t_0 - t_c)^2)$.
(c) $\bar{u}_0(\xi) \geq u_0(\xi)$ and thus $\bar{u}^\epsilon(x,t) \geq u^\epsilon(x,t)$.

From (b), we have

$$\|\bar{u}^\epsilon(\cdot,t_0) - \bar{u}(\cdot,t_0)\| = O\left(\log\frac{1}{|t-t_c|}\right)\epsilon \quad \text{as} \quad \epsilon \to 0, \tag{3.1.19}$$

and from (c)

$$\begin{aligned}
&\|u^\epsilon(\cdot,t_0) - \bar{u}^\epsilon(\cdot,t_0)\|_{L^1(dx;[-1,s(t_0)])} \\
&= \int_{-1}^{s(t_0)} \bar{u}^\epsilon(x,t_0)dx - \int_{-1}^{s(t_0)} u^\epsilon(x,t_0)dx \\
&= -2\epsilon\left[\log\left(\frac{\int e^{-\frac{1}{2\epsilon}\bar{G}(s(t_0),y,t_0)}dy}{\int e^{-\frac{1}{2\epsilon}G(s(t_0),y,t_0)}dy}\right) - \log\left(\frac{\int e^{-\frac{1}{2\epsilon}\bar{G}(-1,y,t_0)}dy}{\int e^{-\frac{1}{2\epsilon}G(-1,y,t_0)}dy}\right)\right],
\end{aligned} \tag{3.1.20}$$

where $\bar{G}(x,y,t) = \int_0^y \bar{u}_0(y')dy' + \frac{(x-y)^2}{2t}$. A standard process of asymptotic expansion leads to

$$\frac{\int e^{-\frac{1}{2\epsilon}\bar{G}(-1,y,t_0)}dy}{\int e^{-\frac{1}{2\epsilon}G(-1,y,t_0)}dy} \sim 1 + O(\epsilon). \tag{3.1.21}$$

As to the first term in (3.1.20), we note that the exponent $G(s(t_0),\cdot,t_0)$ indeed has two global minima occurring at ξ_- and ξ_+ due to the presence of the shock. Here ξ_+ is where the backward characteristic line from $(s(t_0)+0,t_0)$ intersects the x-axis. Since $\bar{u}(\cdot,t_0)$ is smooth, there is only one global minimum of $\bar{G}(s(t_0),\cdot,t_0)$ occurring at ξ_-, therefore,

$$\frac{\int e^{-\frac{1}{2\epsilon}\bar{G}(s(t_0),y,t_0)}dy}{\int e^{-\frac{1}{2\epsilon}G(s(t_0),y,t_0)}dy} \sim \frac{(u_0'(\xi_-)+\frac{1}{t_0})^{\frac{1}{2}}}{(u_0'(\xi_-)+\frac{1}{t_0})^{\frac{1}{2}} + (u_0'(\xi_+)+\frac{1}{t_0})^{\frac{1}{2}}} + o(1) < 1. \tag{3.1.22}$$

From (a), (3.1.19), (3.1.20), (3.1.21), and (3.1.22), we conclude that

$$\|u^\epsilon(\cdot,t_0) - u(\cdot,t_0)\|_{L^1(dx;[-1,s(t_0)])} = O\left(\log\frac{1}{|t-t_c|}\right)\epsilon. \tag{3.1.23}$$

A similar estimate holds for $\|u^\epsilon(\cdot,t_0) - u(\cdot,t_0)\|_{L^1(dx;[s(t_0),1])}$, and the theorem is proved. □

Remark 1. It is clear from the proof that the $O(\epsilon|\log\epsilon|)$ rate is indeed optimal at the critical time. For a general exponent $2p+1$ in (2.3), the constant $\frac{1}{2}$ in (3.1.8) is replaced by $\frac{p}{p+1}$.

The idea used in the proof of Lemma 2.1 and Theorem 3.1 can be carried over to analyze the structure of the error in the case of a centered rarefaction wave. The

$O(\epsilon|\log\epsilon|)$ error is optimal and, roughly speaking, is restricted to the inner edges of the fan.

PROPOSITION 3.3. *Let the initial data $u_0(\xi)$ be a piecewise smooth function with a jump discontinuity at $\xi = 0$ and $u_0(0^-) < u_0(0^+)$ but smooth otherwise. Assume further that u_0 is concave on $\{\xi > 0\}$, convex on $\{\xi < 0\}$, and monotonely nondecreasing. Then the following local L^1 error estimates holds. For c, $d \in R$ satisfying*

$$\frac{u_0(0^+)+u_0(0^-)}{2}t < c < u_0(0^+)t < d,$$

$$\begin{array}{lll} \|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(dx;[\frac{u_0(0^+)+u_0(0^-)}{2}t,c])} & \sim & C_1\epsilon, \\ \|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(dx;[c,u_0(0^+)t])} & \sim & \epsilon|\log\epsilon|, \\ \|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(dx;[u_0(0^+)t,d])} & \sim & C_2\epsilon, \end{array} \tag{3.1.24}$$

where

$$C_1 = 2\log\left(\frac{\frac{1}{\frac{c}{t}-u_0(0^-)}+\frac{1}{u_0(0^+)-\frac{c}{t}}}{\frac{4}{u_0(0^+)-u_0(0^-)}}\right), \quad C_2 = 2\log\left(2\sqrt{\frac{u_0(\xi(u_0(0^+)t,t))}{u_0(\xi(d,t))}}\right), \tag{3.1.25}$$

and $\xi(x,t)$ is defined implicitly by (3.1.3). *Similar estimates hold for intervals at left of the center $\frac{u_0(0_+)+u_0(0_-)}{2}t$.*

The proof is similar to the proof of Theorem 3.1; we omit the detail.

We remark here that the monotonicity and concavity (convexity) assumptions in Proposition 3.3 are not essential; one can treat the case of a general centered rarefaction wave up to the time when the edge is, if ever, merged into a shock. The estimates (3.1.24) remain valid except the constants C_1 and C_2 may become larger due to overestimates. After the edge is merged into a shock, the local L^1 error reduces to $O(\epsilon)$.

The precise form of the statement above is rather complicated; we illustrate with the following example instead.

Example. Consider (2.1) and (2.2) with initial data

$$u_0(\xi) = \begin{cases} -1, & \xi < 0, \\ 1-\frac{\xi}{2}, & 0 \le \xi < 1, \\ -\frac{\xi}{2}, & 1 \le \xi < 2, \\ -1, & 2 \le \xi. \end{cases} \tag{3.1.26}$$

At time $t < 1$, the solution to (1.1) has a rarefaction wave spanning over $-t \le x \le t$ and a standing shock at $\xi = 1$. At $t = 1$, the right edge of the centered rarefaction wave is confronted with the standing shock and merged into it afterward. The following local L^1 estimates hold.

For $0 < t < 1$, we have

$$\begin{array}{lll} \|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(dx;[0,t])} & \sim & \epsilon|\log\epsilon|, \\ \|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(dx;[t,1])} & \sim & O(\epsilon), \\ \|u(\cdot,t)-u^\epsilon(\cdot,t)\|_{L^1(dx;[1,\infty))} & \sim & O(\epsilon). \end{array} \tag{3.1.27}$$

At $t = 1$,

$$\begin{array}{lll} \|u(\cdot,1)-u^\epsilon(\cdot,1)\|_{L^1(dx;[0,1])} & \sim & \epsilon|\log\epsilon|, \\ \|u(\cdot,1)-u^\epsilon(\cdot,1)\|_{L^1(dx;[1,\infty))} & \sim & O(\epsilon). \end{array} \tag{3.1.28}$$

After the interaction, say, $1 < t < 1.5$, the shock begins to move. Denoting by $s(t)$ the shock location, we have

$$\begin{array}{lcl} \|u(\cdot,t) - u^\epsilon(\cdot,t)\|_{L^1(dx;[0,s(t)])} & \sim & O(\epsilon), \\ \|u(\cdot,t) - u^\epsilon(\cdot,t)\|_{L^1(dx;[s(t),\infty))} & \sim & O(\epsilon). \end{array} \tag{3.1.29}$$

We outline the computation for the first equation in (3.1.27); the rest is done in a similar way. Consider

$$\bar{u}_0(\xi) = \begin{cases} -1, & \xi < 0, \\ 1, & \xi > 0, \end{cases} \tag{3.1.30}$$

and denote the corresponding viscous and inviscid solutions by $\bar{u}^\epsilon$ and $\bar{u}$, respectively. By a variant of Lemma 2.1, we can conclude that $\bar{u}(x,t) \geq (\leq) \bar{u}^\epsilon(x,t)$ on $\{x > 0\}$ ($\{x < 0\}$). Thus one can apply the Hopf–Cole transformaton. A short calculation leads to $\bar{G}(x,y,t) = \int_0^y \bar{u}_0(y')dy' + \frac{(x-y)^2}{2t}$ near the absolute minimum,

$$\begin{array}{lcll} \bar{G}(0,y,t) & \sim & |y|, \text{ for } y \text{ near } 0, & \\ \bar{G}(t,y,t) & \sim & \begin{cases} \frac{t}{2} - 2y, \ y < 0 \\ \frac{t}{2} + \frac{y^2}{2t}, \ y > 0 \end{cases} & \text{for } y \text{ near } 0, \end{array} \tag{3.1.31}$$

from which one easily concludes that

$$\|\bar{u}(\cdot,t) - \bar{u}^\epsilon(\cdot,t)\|_{L^1(dx;[0,t])} \sim \epsilon |\log \epsilon|. \tag{3.1.32}$$

On the other hand, $\bar{u}_0 \geq u_0$, thus $\bar{u}^\epsilon \geq u^\epsilon$ and we can apply the Hopf–Cole transformation again. The same calculation leads to

$$\begin{array}{lcll} G(0,y,t) & \sim & |y|, \text{ for } y \text{ near } 0, & \\ G(t,y,t) & \sim & \begin{cases} \frac{t}{2} - 2y, \ y < 0 [1ex] \\ \frac{t}{2} + (\frac{1}{2t} - \frac{1}{4})y^2, \ y > 0 \end{cases} & \text{for } y \text{ near } 0; \end{array} \tag{3.1.33}$$

one thus concludes that

$$\|\bar{u}^\epsilon(\cdot,t) - u^\epsilon(\cdot,t)\|_{L^1(dx;[0,t])} \sim \epsilon \log\left(\frac{2-t}{2}\right). \tag{3.1.34}$$

We conclude with the first equation of (3.1.27), with the triangle inequality and the fact that $u(x,t)$ coincides with $\bar{u}(x,t)$ for $x \leq t$.

3.2. Hopf–Cole–Lax transformation for some monotone schemes. We now give another example in which the convergence rate is not first order at the critical time—Lax–Friedrichs scheme applied to the conservation law (1.1) with a specific flux function:

$$f_L(u) = \log\left(\frac{\cosh(u)+1}{2}\right).$$

The Lax–Friedrichs scheme for this particular flux admits a discrete version of Hopf–Cole transformation. This was first observed by Lax [7] for upwind scheme with a family of flux function $f(u) = \log(a + be^{-u})$, $a, b > 0$, $a + b = 1$. Here we adopt a variation of the original one in order to maintain symmetry, which simplifies the analysis.

The following properties of $f_L(u)$ are elementary:

(C.1) $f_L(u) = f_L(-u)$.
(C.2) $f_L(0) = 0$.
(C.3) $f_L'(u) = \frac{\sinh(u)}{\cosh(u)+1}$, $f_L'(0) = 0$.
(C.4) $f_L''(u) = \frac{1}{\cosh(u)+1} > 0$.

We now study the convergence rate for the Lax–Friedrichs scheme with f_L. Let $u^\Delta(x,t)$ be the approximate solution obtained via the Lax–Friedrichs scheme

$$u^\Delta(x,t+\Delta) = \frac{1}{2}(\, u^\Delta(x-\Delta,t)+u^\Delta(x+\Delta,t)\,) - \frac{1}{2}(\, f_L(u^\Delta(x+\Delta,t)) - f_L(u^\Delta(x-\Delta,t))\,), \tag{3.2.1}$$

where the argument (x,t) is restricted to grid points only, and we have put $\Delta x = \Delta t = \Delta$ for simplicity.

Now let $U^\Delta(x,t) = 2\Delta \sum_{k=-\infty^0} u^\Delta(x-2k\Delta,t)$. The equation for U^Δ is

$$U^\Delta(x,t+\Delta) = \frac{1}{2}(\, U^\Delta(x-\Delta,t)+U^\Delta(x+\Delta,t)\,) - \Delta f_L\left(\frac{U^\Delta(x+\Delta,t) - U^\Delta(x-\Delta,t)}{2\Delta}\right). \tag{3.2.2}$$

Now we apply the Hopf–Cole–Lax transformation

$$U^\Delta = G(V^\Delta) = -2\Delta \log(V^\Delta),$$

which brings (3.2.1) to

$$\begin{aligned} G(V^\Delta(x,t+\Delta)) = \quad & \frac{1}{2}[G(V^\Delta(x+\Delta,t)) + G(V^\Delta(x-\Delta,t))][1ex] \\ & -\Delta f_L\left(\frac{G(V^\Delta(x+\Delta,t)) - G(V^\Delta(x-\Delta,t))}{2\Delta}\right). \end{aligned} \tag{3.2.3}$$

The equation for V^Δ thus linearizes as the following identity holds for all V, $W \in R$,

$$\frac{1}{2}(G(V)+G(W)) - \Delta f_L\left(\frac{G(V)-G(W)}{2\Delta}\right) = G\left(\frac{V+W}{2}\right).$$

Thus

$$V^\Delta(x,t+\Delta) = \frac{1}{2}(\, V^\Delta(x+\Delta,t) + V^\Delta(x-\Delta,t)\,),$$

and, therefore,

$$V^\Delta(x,t) = \sum_{l=0}^{n} \binom{n}{l} \frac{1}{2^n} V^\Delta(x-n\Delta+2\Delta,0),$$

where $t = n\Delta$.

For fixed $x, z \in \mathbf{R}$, $t > 0$, we want to estimate

$$U^\Delta(x,t) - U^\Delta(z,t) = -2\Delta \log\left(\frac{\displaystyle\sum_{l=0}^{n}\binom{n}{l} e^{-\frac{1}{\Delta}U^\Delta(x-n\Delta+2l\Delta,0)}}{\displaystyle\sum_{l=0}^{n}\binom{n}{l} e^{-\frac{1}{\Delta}U^\Delta(z-n\Delta+2l\Delta,0)}}\right), \tag{3.2.4}$$

where we've used $U^\Delta(\cdot,0) = -2\Delta\log(V^\Delta(\cdot,0))$. For the sake of a simpler formula, we assume that (x,t) and (z,t) are always on the grid points as the mesh refines.

The following counterpart of Lemma 2.1 is crucial in establishing the ordering of $u^\Delta(\cdot,t_c)$ and $u(\cdot,t_c)$; therefore, we can estimate the L^1 difference of the two using (3.2.4).

LEMMA 3.4. *Let u_0 be a smooth and bounded function satisfying*

(B.1) $f_L'(u_0(\xi)) = -\frac{\xi}{t_c} + a\xi^3 + b\xi^4 + O(\xi^5)$ *for* $|\xi| < \delta$ *where* $a > 0$.

(B.2) $\xi = 0$ *is the point corresponding to the first spontaneous formation of shocks; that is,* $\frac{d}{d\xi} f_L'(u_0(\xi)) > -\frac{1}{t_c}$ *for all* $\xi \neq 0$.

(B.3) $f_L'(u_0)$ *is antisymmetric, and thus so is* u_0*:* $u_0(-\xi) = -u_0(\xi)$.

(B.4) $f_L'(u_0(\cdot))$*, and, therefore,* $u_0(\cdot)$ *is monotonely nonincreasing.*

(B.5) $f_L'(u_0(\cdot))$ *is concave on* $\xi < 0$*, and, therefore, by Assumption* (B.3)*, convex on* $\xi > 0$.

Then

$$u(x,\Delta) \geq u^\Delta(x,\Delta) \quad x < 0,\ (x,\Delta) \text{ on the grids.}$$

By induction and the monotonicity of Lax–Friedrichs scheme, $u(x,t) \geq u^\Delta(x,t)$ *for all* $x < 0$, $t > 0$, (x,t) *on the grids.*

Proof. Let $A = (x,\Delta), B = (x-\Delta,0), C = (x+\Delta,0)$, and $D = (x,0)$ be four-grid points on $x < 0$; then

- $u_A^\Delta = g(u_B,u_C) \equiv \frac{1}{2}(u_B + u_C) - \frac{1}{2}(f_L(u_C) - f_L(u_B))$;
- $u_B \geq u_C$, and $u_A = u_E$ for some (unique) point E on the line segment $\bar{BD}$.

Denote by $m = f_L'(u_E) = \mathit{distance}(D,E)/\Delta$, and define, for $0 \leq \theta \leq 1$, a family of functions $v(\theta,\xi)$ on $\bar{BC}$ by

$$f_L'(v(\theta,\xi)) = m + \theta(f_L'(u_0(\xi)) - m), \quad 0 \leq \theta \leq 1,\ x-\Delta \leq \xi \leq x+\Delta.$$

Obviously, $v(0,\xi) = u_E$ and $v(1,\xi) = u_0(\xi)$.

Now let $h(\theta) = g(v(\theta,x-\Delta), v(\theta,x+\Delta))$; then $h(0) = u_E$ and $h(1) = g(u_B,u_C) = u_A^\Delta$. A direct computation gives

$$\frac{dh}{d\theta} = \frac{1}{2}(\alpha + \theta\alpha\beta + m\beta),$$

where $\alpha = f'(u_B) + f'(u_C) - 2m < 0$ and $\beta = f'(u_B) - f'(u_C) > 0$. Due to concavity of $f'(u_0(\cdot))$, the graph of $f'(u_0(\cdot))$ lies above the line joining $(B, f'(u_B))$ and $(C, f'(u_C))$; therefore, $\alpha + m\beta \leq 0$. Thus $\frac{dh}{d\theta} \leq 0$ and $u_A \geq u_A^\Delta$. ∎

Now we come back to estimate the leading order term of (3.2.4) using Stirling's formula

$$n! = \left(\frac{n-1}{e}\right)^{n-1} (2\pi(n-1))^{\frac{1}{2}} + \cdots. \tag{3.2.5}$$

After elementary calculations, we have

$$U^\Delta(x,t) - U^\Delta(z,t) = -2\Delta\log\left(\frac{\sum_{l=0}^{n} e^{-\frac{1}{\Delta}[(tF(\frac{x-y}{t}) + U_0(y,o)) + E_1]}}{\sum_{l=0}^{n} e^{-\frac{1}{\Delta}[(tF(\frac{z-y}{t}) + U_0(y,o)) + E_2]}}\right), \tag{3.2.6}$$

where $F(s) = \log(1-s^2) + s\log(\frac{1+s}{1-s})$, $t = n\Delta$, and $x - y = t - 2l\Delta$. E_1 and E_2 are the errors introduced by Stirling's formula, and are of lower order.

We next replace sums by integrals. Again, the errors are of lower order since the integrals are at least $O(\Delta^{\frac{1}{2}})$ as we saw in section 3.1. This leads to

$$U^\Delta(x,t) - U^\Delta(z,t) = -2\Delta \log \left(\frac{\int_{x-t}^{x+t} e^{-\frac{1}{2\Delta}[(tF(\frac{x-y}{t}) + \int^y u_0(y')dy') + E_1]} dy}{\int_{z-t}^{z+t} e^{-\frac{1}{2\Delta}[(tF(\frac{z-y}{t}) + \int^y u_0(y')dy') + E_2]} dy} \right) + \cdots. \tag{3.2.7}$$

At $x = 0$, $t = t_c$, the integrand of the numerator has a quartic phase at its maximum $y = 0$, while the integrand of the denominator has a quadratic phase at $z =$, say, -1, $t = t_c$. Therefore,

$$U^\Delta(0,t_c) - U^\Delta(-1,t_c) = \int_{-1}^{0} u(.,t_c) - 2\Delta \log \left(\frac{\Delta^{\frac{1}{4}}}{\Delta^{\frac{1}{2}}} \right) + \cdots,$$

and

$$\|u(\cdot,t_c) - u^\Delta(\cdot,t_c)\|_{L^1(\Delta x;[-1,0])} \sim \frac{1}{2}\Delta|\log\Delta|.$$

The generalization to nonantisymmetric initial data is the same as for Burgers's equation in the previous subsection.

Remark 2.

1. The case $t \neq t_c$ can be proved in the same way; see [2] for a discrete version of Lemma 2.2. The discrete analogue of the comparison lemma is an immediate consequence of monotonicity.

2. Although we only carry out the analysis for the most dissipative first-order scheme, namely, the Lax–Friedrichs scheme, the same argument shows that even the upwind scheme cannot do better. Since the same transform applies to the upwind scheme with the flux function $f(u) = -\log(a+be^{-u})$, $a, b > 0, a+b = 1$. Even though we don't have symmetry in this case, we still have the lower bound for free:

$$\|u(\cdot,t_c) - u^\Delta(\cdot,t_c)\|_{L^1(\Delta x;[x_c-1,x_c])} \geq |U(x_c,t_c) - U^\Delta(x_c-1,t_c)| = O(\Delta|\log\Delta|) \tag{3.2.8}$$

3. Since $f'(u) = \frac{be^{-u}}{a+be^{-u}} > 0$, the Godunov scheme reduces to upwind scheme. Therefore, (3.2.8) also holds for Godunov scheme with the same family of flux.

Acknowledgments. The author would like to thank his advisor, Professor Z. Xin, for suggesting this problem and some helpful discussions. He would also like to thank Professor J. Goodman and Dr. S. H. Yu for reminding him of the discrete Hopf–Cole transformation in [7].

REFERENCES

[1] J. M. Hyman, A. Harten, and P. D. Lax, *On finite-difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), pp. 297–322.

[2] M. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[3] B. Engquist and S.H. Yu, *Convergence of Finite Difference Schemes for Piecewise Smooth Solutions with Shocks*, preprint, 1995.
[4] J. Goodman and Z. Xin, *Viscous limits for piecewise smooth solutions to systems of conservation laws*, Arch. Rationál Mech. Anal., 12 (1992), pp. 235–265.
[5] E. Harabetian, *Rarefactions and large time behavior for parabolic equations and monotone schemes*, Comm. Math. Phys., 114 (1988), pp. 527–536.
[6] N. N. Kuznetsov, *Accuracy of some approximate methods for scalar scalar conservation laws*, USSR Comput. Math. Math. Phys., 16 (1976), pp. 105–119.
[7] P. D. Lax, *Hyperbolic systems of conservation laws* (II), Comm. Pure Appl. Math., 10 (1957), pp. 537–566.
[8] B. J. Lucier, *Error bounds for the methods of Glimm, Godunov and LeVeque*, SIAM J. Numer. Anal., 22 (1985), pp. 1074–1081.
[9] F. Sabac, *The Optimal Convergence Rate of Monotone Finite Difference Schemes*, Ph.D. thesis, University of South Carolina, Columbia, SC, 1995.
[10] R. Sanders, *On convergence of monotone finite difference schemes with variable spatial differencing*, Math. Comp., 40 (1983), pp. 91–106.
[11] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1983.
[12] Y. Smyrlis and S. H. Yu, *Existence and stability of traveling discrete shocks: Numerical analysis and its applications* (*Rousse*, 1996), Lecture Notes in Comput. Sci. 1196, Springer, Berlin, 1997, pp. 466–473.
[13] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.
[14] T. Tang and Z. H. Teng, *The sharpness of Kuznetsov's* $O(\sqrt{\Delta x})$ L^1 *error estimate for monotone difference schemes*, Math. Comp., 64 (1995), pp. 581–589.
[15] T. Tang and Z. H. Teng, *Viscosity methods for piecewise smooth solutions to scalar conservation laws*, Math. Comp., 66 (1997), pp. 495–526.
[16] Z.- H. Teng and P. Zhang, *Optimal* L^1 *Rate of convergence for the viscosity method and monotone scheme to piecewise constant solutions with shocks*, SIAM J. Numer. Anal., 34 (1997), pp. 959–978.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 53–71

ON A WAVE EQUATION WITH A BOUNDARY CONDITION ASSOCIATED WITH CAPILLARY WAVES*

JONG UHN KIM†

Abstract. This paper discusses an initial-boundary value problem for a wave equation with a nonstandard boundary condition associated with linear capillary waves on the surface of a compressible liquid. We prove the well-posedness of this problem. Our main technical device is the Fourier transform.

Key words. wave equation, initial-boundary value problem, capillary waves

AMS subject classifications. 35B30, 35D05, 35L05

PII. S003614109732821X

Introduction. In this paper we discuss the following initial-boundary value problem for a wave equation:

$$u_{tt} - \Delta u = 0 \qquad \text{for } t \in (0,T),\ (x,y,z) \in R^3_+, \tag{0.1}$$

$$\mathcal{B}u(t,x,y,0) = 0 \qquad \text{for } t \in (0,T),\ (x,y) \in R^2, \tag{0.2}$$

$$u(0,x,y,z) = u_0(x,y,z),\ \ u_t(0,x,y,z) = u_1(x,y,z) \quad \text{for } (x,y,z) \in R^3_+, \tag{0.3}$$

where $T > 0$ is arbitrarily given, $R^3_+ = \{(x,y,z) : (x,y) \in R^2,\ z > 0\}$, Δ is the Laplacian in (x,y,z), and the operator $\mathcal{B}$ is defined by

$$\mathcal{B}u = \Delta u + \frac{\partial}{\partial z}\big(u_{xx} + u_{yy}\big). \tag{0.4}$$

This nonstandard boundary condition arises in the linearized model of capillary waves. Let a compressible liquid occupy the half-space R^3_+, and let u denote its velocity potential. When the effects of gravity are neglected, u satisfies the following acoustic equation within the liquid:

$$u_{tt} - c^2\Delta u = 0, \tag{0.5}$$

where c is the speed of sound. We describe the free surface by $z = \zeta(t,x,y)$. By the Laplace formula (see Landau and Lifshitz [5]), we have

$$p - p_0 = \alpha\big(\zeta_{xx} + \zeta_{yy}\big), \tag{0.6}$$

where $\alpha > 0$ is the surface-tension coefficient, p is the pressure in the liquid near the surface, and p_0 is the constant external pressure. On the other hand, u is related to the pressure by

$$p - p_0 = -\rho_0 \frac{\partial u}{\partial t}, \tag{0.7}$$

*Received by the editors October 3, 1997; accepted for publication (in revised form) March 4, 1998; published electronically September 29, 1998.
http://www.siam.org/journals/sima/30-1/32821.html
†Department of Mathematics, Virginia Tech, Blacksburg, VA 24061-0123 (kim@math.vt.edu).

where ρ_0 is the constant equilibrium density and gravity has been neglected. On the surface, it holds that

$$\frac{\partial u}{\partial z} = \frac{\partial \zeta}{\partial t} \qquad \text{at } z = 0. \tag{0.8}$$

Combining (0.5)–(0.8), we have

$$\Delta u + \frac{\alpha}{\rho_0 c^2} \frac{\partial}{\partial z}\big(u_{xx} + u_{yy}\big) = 0 \qquad \text{at } z = 0. \tag{0.9}$$

By rescaling the variables, we arrive at (0.1) and (0.2).

For the Dirichlet boundary condition $u = 0$, or the Neumann boundary condition $\frac{\partial u}{\partial z} = 0$ at $z = 0$, a complete result on the well-posedness of the initial-boundary value problem is well known. The purpose of this work is to establish a similar result in the L^2 Sobolev spaces with the boundary condition (0.2). Bondi [2] constructed a fundamental solution of (0.1)–(0.2) and studied an interesting phenomenon of wave propagation along the surface with infinite speed. His result was also discussed by Duff [3]. Here our issue is to formulate the well-posedness of the initial-boundary value problem (0.1)–(0.3) as closely as possible to that of the standard initial-boundary value problems. One could use fundamental solutions to represent solutions. However, even for the pure Cauchy problem in the whole space, the explicit formula of the fundamental solution is not a good device to obtain the basic L^2 energy estimates, which trivially follow either by a multiplier technique or by the Fourier transform. In particular, it is not at all clear whether the fundamental solution obtained in [2] can be used to obtain estimates of solutions in the function classes of L^2 setting. Here we rely on the well-known theory of the Cauchy problem in the whole space. This involves a suitable extension of the given data to the whole space, where various compatibility conditions are naturally incorporated. The main tool is the Fourier transform. When the boundary operator consists of only odd (or only even) derivatives in the normal coordinate variable, this is a standard procedure. The novelty of the above problem is that the boundary operator is hybrid, and a simple reflection method does not work. So the main difficulty lies in handling the boundary condition. The boundary $z = 0$ is noncharacteristic, and the trace of any derivative of the solution is well-defined due to partial hypoellipticity; see Hörmander [4]. But we need some regularity condition for the uniqueness of the solution, which is justified by a simple example presented in the next section. Our starting point is the following observation made by Duff [3]. Since $\mathcal{B}$ is a differential operator with constant coefficients, $\mathcal{B}u$ satisfies the wave equation if u does. Consequently, $\mathcal{B}u$ is a solution of the initial-boundary value problem with the initial data $(\mathcal{B}u_0, \mathcal{B}u_1)$ and the Dirichlet boundary condition. However, it is only a weak solution unless u itself is sufficiently smooth. For this, we provide in the next section a definition of a weak solution which is suitable in our function classes such that each solution in the natural energy space is included. As the initial data are more regular with suitable compatibility conditions at the boundary $z = 0$, the regularity of the solution improves accordingly.

Following the notation for function classes in Lions and Magenes [6], we state the main result as follows.

THEOREM 0.1. *Let $u_0 \in H^2(R^3_+)$ and $u_1 \in H^1(R^3_+)$ with $\frac{\partial u_0}{\partial z}(x, y, 0) \in H^1(R^2)$. Then, there is a unique solution $u \in C([0,T]\,;\, H^2(R^3_+)) \cap C^1([0,T]\,;\, H^1(R^3_+))$ of (0.1)–(0.3) such that $\mathcal{B}u$ is a solution of (1.9) below with the initial data $(\mathcal{B}u_0, \mathcal{B}u_1)$.*

Furthermore, for each $t \in [0,T]$ the solution satisfies

$$\begin{aligned}(0.10)\qquad &\|u(t,\cdot)\|_{H^2(R^3_+)} + \|u_t(t,\cdot)\|_{H^1(R^3_+)}\\ &\quad\le M(T)\bigg(\|u_0\|_{H^2(R^3_+)} + \|u_1\|_{H^1(R^3_+)} + \Big\|\frac{\partial u_0}{\partial z}(\cdot,0)\Big\|_{H^1(R^2)}\bigg),\end{aligned}$$

where $M(T)$ is a positive constant depending on T.

We can also obtain more regular solutions.

COROLLARY 0.2. *If we suppose that $u_0 \in H^3(R^3_+)$ with $\mathcal{B}u_0 = 0$ at $z = 0$ and that $u_1 \in H^2(R^3_+)$ with $\frac{\partial u_1}{\partial z}(\cdot,0) \in H^1(R^2)$, the above solution belongs to $C([0,T]\,;\, H^3(R^3_+)) \cap C^1([0,T]\,;\, H^2(R^3_+))$, and it holds that for all $t \in [0,T]$,*

$$\begin{aligned}(0.11)\qquad &\|u(t,\cdot)\|_{H^3(R^3_+)} + \|u_t(t,\cdot)\|_{H^2(R^3_+)}\\ &\quad\le M(T)\bigg(\|u_0\|_{H^3(R^3_+)} + \|u_1\|_{H^2(R^3_+)} + \Big\|\frac{\partial u_1}{\partial z}(\cdot,0)\Big\|_{H^1(R^2)}\bigg)\end{aligned}$$

for some positive constant $M(T)$.

COROLLARY 0.3. *Let $m \ge 1$ be an integer. Suppose that $u_0 \in H^{2m+2}(R^3_+)$, $\mathcal{B}u_0 = 0$, $\Delta\mathcal{B}u_0 = 0, \dots, \Delta^{m-1}\mathcal{B}u_0 = 0$ at $z = 0$, $(\frac{\partial}{\partial z})^{2m+1}u_0(\cdot,0) \in H^1(R^2)$ and that $u_1 \in H^{2m+1}(R^3_+)$, $\mathcal{B}u_1 = 0$, $\Delta\mathcal{B}u_1 = 0, \dots, \Delta^{m-1}\mathcal{B}u_1 = 0$ at $z = 0$. Then, the above solution belongs to $C([0,T]\,;\, H^{2m+2}(R^3_+)) \cap C^1([0,T]\,;\, H^{2m+1}(R^3_+))$ and satisfies for all $t \in [0,T]$,*

$$\begin{aligned}(0.12)\quad &\|u(t,\cdot)\|_{H^{2m+2}(R^3_+)} + \|u_t(t,\cdot)\|_{H^{2m+1}(R^3_+)}\\ &\quad\le M(T,m)\bigg(\|u_0\|_{H^{2m+2}(R^3_+)} + \|u_1\|_{H^{2m+1}(R^3_+)} + \Big\|\frac{\partial^{2m+1}u_0}{\partial z^{2m+1}}(\cdot,0)\Big\|_{H^1(R^2)}\bigg),\end{aligned}$$

where $M(T,m)$ is a positive constant depending on T and m.

COROLLARY 0.4. *Let $m \ge 1$ be an integer. Suppose that $u_0 \in H^{2m+3}(R^3_+)$, $\mathcal{B}u_0 = 0$, $\Delta\mathcal{B}u_0 = 0, \dots, \Delta^m\mathcal{B}u_0 = 0$ at $z = 0$ and that $u_1 \in H^{2m+2}(R^3_+)$, $\mathcal{B}u_1 = 0$, $\Delta\mathcal{B}u_1 = 0, \dots, \Delta^{m-1}\mathcal{B}u_1 = 0$ at $z = 0$, $(\frac{\partial}{\partial z})^{2m+1}u_1(\cdot,0) \in H^1(R^2)$. Then, the above solution belongs to $C([0,T]\,;\, H^{2m+3}(R^3_+)) \cap C^1\big([0,T]\,;\, H^{2m+2}(R^3_+)\big)$ and satisfies for all $t \in [0,T]$,*

$$\begin{aligned}(0.13)\quad &\|u(t,\cdot)\|_{H^{2m+3}(R^3_+)} + \|u_t(t,\cdot)\|_{H^{2m+2}(R^3_+)}\\ &\quad\le M(T,m)\bigg(\|u_0\|_{H^{2m+3}(R^3_+)} + \|u_1\|_{H^{2m+2}(R^3_+)} + \Big\|\frac{\partial^{2m+1}u_1}{\partial z^{2m+1}}(\cdot,0)\Big\|_{H^1(R^2)}\bigg).\end{aligned}$$

If we assume that the support of the initial data is contained in R^3_+, then the result is essentially the same as that for the standard boundary conditions.

COROLLARY 0.5. *Let m be a nonnegative integer. Let $u_0 \in H^{m+2}(R^3_+)$ and $u_1 \in H^{m+1}(R^3_+)$ such that $\operatorname{supp} u_0 \cup \operatorname{supp} u_1 \subset R^3_+$. Then, the above solution belongs to $C([0,T]\,;\, H^{m+2}(R^3_+)) \cap C^1([0,T]\,;\, H^{m+1}(R^3_+))$ and satisfies for each $t \in [0,T]$,*

$$\begin{aligned}(0.14)\qquad &\|u(t,\cdot)\|_{H^{m+2}(R^3_+)} + \|u_t(t,\cdot)\|_{H^{m+1}(R^3_+)}\\ &\quad\le M(T,m)(\|u_0\|_{H^{m+2}(R^3_+)} + \|u_1\|_{H^{m+1}(R^3_+)}).\end{aligned}$$

We will prove these results in the next sections.

1. Preliminaries. We first review some facts about the Cauchy problem for a wave equation in R^3, which is formulated as follows:

$$\begin{cases} u_{tt} - \Delta u = 0 \qquad \text{in } (0,T)\times R^3, \\ u(0,x,y,z) = u_0(x,y,z) \qquad \text{in } R^3, \\ u_t(0,x,y,z) = u_1(x,y,z) \qquad \text{in } R^3, \end{cases} \tag{1.1}$$

where Δ is the Laplacian in $(x,y,z)\in R^3$, and $0<T<\infty$ is arbitrarily given.

LEMMA 1.1. *Let $u_0 \in H^s(R^3)$, and $u_1 \in H^{s-1}(R^3)$, for some $s \in R$. Then, there is a unique solution of the Cauchy problem* (1.1) *in $C([0,T]\,;\,H^s(R^3))\cap C^1([0,T]\,;\,H^{s-1}(R^3))$, and it holds that*

$$\|u(t,\cdot)\|_{H^s(R^3)} + \|u_t(t,\cdot)\|_{H^{s-1}(R^3)} \le M(T)(\|u_0\|_{H^s(R^3)} + \|u_1\|_{H^{s-1}(R^3)}) \tag{1.2}$$

for all $t\in[0,T]$ with some positive constant $M(T)$ independent of u_0 and u_1.

LEMMA 1.2. *In the same setting as above, assume that* supp $u_0 \cup$ supp $u_1 \subset \{(x,y,z) : z < -L\}$ *for some $L\in R$. Then, the support of the solution for $t\ge 0$ is contained in $\{(t,x,y,z) : t > z+L\}$.*

Next we define a function space $\mathcal{X}$ by

$$\begin{aligned} \mathcal{X} =\ & H^{-1}(R_z\,;\,H^{-1}(R^2)) + L^2(R_z\,;\,H^{-2}(R^2)) \\ & = \{f_1+f_2 \,:\, f_1(x,y,z)\in H^{-1}(R_z\,;\,H^{-1}(R^2)),\ \ f_2(x,y,z)\in L^2(R_z\,;\,H^{-2}(R^2))\}, \end{aligned}$$

equipped with the norm

$$\|f\|_{\mathcal{X}} = \inf_{f=f_1+f_2} (\|f_1\|_{H^{-1}(R_z;H^{-1}(R^2))} + \|f_2\|_{L^2(R_z;H^{-2}(R^2))}), \tag{1.3}$$

where R_z is for the z variable. Then, $\mathcal{X}$ is the dual of $H^1(R_z\,;\,H^1(R^2))\cap L^2(R_z\,;\,H^2(R^2))$. For this, see Bergh and Löfström [1]. We also need the following fact.

LEMMA 1.3. *The operator $(1-\Delta_{x,y})^{1/2}$ is an isomorphism from $H^{-1}(R^3)$ onto $\mathcal{X}$. Here, $\Delta_{x,y}$ is the Laplacian in (x,y).*

Proof. Choose any $f\in H^{-1}(R^3)$, and set

$$\hat{f}_1(\xi) = \frac{|\xi_3|}{1+|\xi_1|+|\xi_2|+|\xi_3|}\hat{f}(\xi)$$

and

$$\hat{f}_2(\xi) = \frac{1+|\xi_1|+|\xi_2|}{1+|\xi_1|+|\xi_2|+|\xi_3|}\hat{f}(\xi),$$

where $\hat{f}(\xi)$ is the Fourier transform of f and $\xi=(\xi_1,\xi_2,\xi_3)$ is the dual variable of (x,y,z). It is apparent that $f=f_1+f_2$ and

$$\|(1-\Delta_{x,y})^{1/2} f_1\|_{H^{-1}(R_z;H^{-1}(R^2))} \le M\|f\|_{H^{-1}(R^3)} \tag{1.4}$$

and

$$\|(1-\Delta_{x,y})^{1/2} f_2\|_{L^2(R_z;H^{-2}(R^2))} \le M\|f\|_{H^{-1}(R^3)} \tag{1.5}$$

for some positive constant M independent of f. In the meantime, $H^{-1}(R_z\,;\,L^2(R^2))$ and $L^2(R_z\,;\,H^{-1}(R^2))$ are continuously embedded into $H^{-1}(R^3)$. Hence, $(1-\Delta_{x,y})^{-1/2}$ is continuous from $\mathcal{X}$ into $H^{-1}(R^3)$. □

LEMMA 1.4. *Suppose that* $u_0 \in L^2(R_z\,;\,H^{-1}(R^2))$ *and* $u_1 \in \mathcal{X}$. *Then, there is a unique solution* u *of* (1.1) *in* $C\big([0,T]\,;\,L^2(R_z\,;\,H^{-1}(R^2))\big) \cap C^1([0,T]\,;\,\mathcal{X})$. *We also have*

$$\|u(t,\cdot)\|_{L^2(R_z;H^{-1}(R^2))} + \|u_t(t,\cdot)\|_{\mathcal{X}} \le M(T)\left(\|u_0\|_{L^2(R_z;H^{-1}(R^2))} + \|u_1\|_{\mathcal{X}}\right) \tag{1.6}$$

for all $t \in [0,T]$, *where* $M(T)$ *is a positive constant independent of* u_0 *and* u_1.

Proof. Let us set

$$v_0 = (1-\Delta_{x,y})^{-1/2}u_0, \qquad v_1 = (1-\Delta_{x,y})^{-1/2}u_1, \tag{1.7}$$

and consider the Cauchy problem (1.1) with (u_0,u_1) replaced by (v_0,v_1). Since $v_0 \in L^2(R^3)$ and $v_1 \in H^{-1}(R^3)$, it follows from Lemma 1.1 that there is a unique solution v in $C([0,T]\,;\,L^2(R^3)) \cap C^1([0,T]\,;\,H^{-1}(R^3))$. We then set

$$u = (1-\Delta_{x,y})^{1/2}v. \tag{1.8}$$

By virtue of Lemmas 1.1 and 1.3, we find that u is the unique solution of (1.1) in $C([0,T]\,;\,L^2(R_z\,;\,H^{-1}(R^2))) \cap C^1([0,T]\,;\mathcal{X})$ and that (1.6) is satisfied.

We now consider an initial-boundary value problem in a half-space:

$$\begin{cases} u_{tt} - \Delta u = 0 \qquad \text{in} \quad (0,T)\times R^3_+, \\ u(t,x,y,0) = 0 \qquad \text{in} \quad (0,T)\times R^2, \\ u(0,x,y,z) = u_0(x,y,z) \qquad \text{in} \quad R^3_+, \\ u_t(0,x,y,z) = u_1(x,y,z) \qquad \text{in} \quad R^3_+. \end{cases} \tag{1.9}$$

We define a function space $\mathcal{X}_1$ by

$$\mathcal{X}_1 = \{f_1 + f_2 \,:\, f_1 \in H^{-1}((0,\infty)\,;\,H^{-1}(R^2)),\ f_2 \in L^2(0,\infty\,;\,H^{-2}(R^2))\}$$

equipped with the norm

$$\|f\|_{\mathcal{X}_1} = \inf_{f=f_1+f_2}\left(\|f_1\|_{H^{-1}((0,\infty);H^{-1}(R^2))} + \|f_2\|_{L^2(0,\infty;H^{-2}(R^2))}\right), \tag{1.10}$$

so that $\mathcal{X}_1$ is the dual of $H^1_0((0,\infty)\,;\,H^1(R^2)) \cap L^2(0,\infty\,;\,H^2(R^2))$. We then adopt the following definition of a solution of (1.9).

DEFINITION 1.5. *Let* $u_0 \in L^2(0,\infty\,;\,H^{-1}(R^2))$ *and* $u_1 \in \mathcal{X}_1$. *We say that a function* $u \in C\big([0,T]\,;\,L^2(0,\infty;\,H^{-1}(R^2))\big) \cap C^1([0,T]\,;\,\mathcal{X}_1)$ *is a solution of* (1.9) *if it holds that*

$$\begin{aligned} -\langle u_1,\,\phi(0,x,y,z)\rangle_1 + \langle u_0,\,\phi_t(0,x,y,z)\rangle_2 \\ + \int_0^\infty \int_0^T \langle u,\ \phi_{tt} - \Delta\phi\rangle_3 \,dt\,dz = 0 \end{aligned} \tag{1.11}$$

for every $\phi \in C^2([0,T]\,;\,H^2(R^3_+))$ *such that*

$$\begin{cases} \phi(t,x,y,0) = 0 \qquad \textit{in} \quad [0,T]\times R^2, \\ \phi(T,x,y,z) = 0, \qquad \phi_t(T,x,y,z) = 0 \qquad \textit{in} \quad R^3_+. \end{cases} \tag{1.12}$$

Here, the bracket $\langle\cdot,\cdot\rangle_1$ *is the duality pairing between* $\mathcal{X}_1$ *and* $H^1_0((0,\infty)\,;\,H^1(R^2)) \cap L^2(0,\infty\,;\,H^2(R^2))$, $\langle\cdot,\cdot\rangle_2$ *is the duality pairing between* $L^2(0,\infty\,;\,H^{-1}(R^2))$ *and* $L^2(0,\infty\,;\,H^1(R^2))$, *and* $\langle\cdot,\cdot\rangle_3$ *is the duality pairing between* $H^{-1}(R^2)$ *and* $H^1(R^2)$.

First we note that the solution in the natural energy space $C([0,T]\,;\,H_0^1(R^3_+))\cap C^1([0,T]\,;\,L^2(R^3_+))$ satisfies the above condition (1.11). We will show that the condition (1.11) implies the boundary condition at $z=0$. Choose any $\psi\in C_0^\infty((0,T)\times R^2)$, and set

$$J(z)=\int_0^T \langle u(t,x,y,z),\,\psi(t,x,y)\rangle_3\,dt,$$

$$I(z)=\int_0^T \langle u(t,x,y,z),\;\psi_{tt}-\Delta_{x,y}\psi\rangle_3\,dt.$$

Then, $J(z)$ and $I(z)$ belong to $L^2((0,\infty))$, and it holds that

$$\frac{d^2J}{dz^2}(z)=I(z),\qquad \text{in}\quad \mathcal{D}'((0,\infty)), \tag{1.13}$$

since (1.11) implies that u satisfies the wave equation in $\mathcal{D}'((0,T)\times R^3_+)$. Thus, $J(z)\in H^2((0,\infty))\subset C^1([0,T])$. Next choose $p(z)\in C_0^\infty([0,\infty))$ with $p(0)=0$, and $\frac{dp}{dz}(0)=1$. It follows from (1.13) that

$$\int_0^\infty \frac{d^2J}{dz^2}(z)\,p(z)\,dz=\int_0^\infty I(z)p(z)\,dz. \tag{1.14}$$

But we have

$$\int_0^\infty \frac{d^2J}{dz^2}\,p(z)\,dz=J(0)+\int_0^\infty J(z)\,\frac{d^2p}{dz^2}(z)\,dz. \tag{1.15}$$

By (1.11), we find that

$$\begin{aligned}0&=\int_0^\infty\int_0^T \langle u,\psi_{tt}-\Delta_{x,y}\psi\rangle_3\,p(z)\,dt\,dz-\int_0^\infty\int_0^T\langle u,\psi\rangle_3\,\frac{d^2p}{dz^2}(z)\,dt\,dz\\ &=\int_0^\infty I(z)p(z)\,dz-\int_0^\infty J(z)\frac{d^2p}{dz^2}(z)\,dz.\end{aligned} \tag{1.16}$$

It follows from (1.14)–(1.16) that $J(0)=0$, which means that u satisfies the boundary condition. Next we show that (1.11) implies that the initial conditions are satisfied. Choose an arbitrary function $\theta(x,y,z)\in C_0^\infty(R^3_+)$, and set

$$K(t)=\int_0^\infty \langle u(t,x,y,z),\,\theta(x,y,z)\rangle_3\,dz,$$

$$L(t)=\int_0^\infty \langle u(t,x,y,z),\,\Delta\theta(x,y,z)\rangle_3\,dz.$$

Then, $K(t)$ and $L(t)$ are continuous in $t\in[0,T]$, and

$$\frac{d^2K}{dt^2}(t)=L(t)\qquad \text{holds in}\quad \mathcal{D}'((0,T)), \tag{1.17}$$

which yields $K(t) \in C^2([0,T])$. Next choose a function $q(t) \in C^\infty([0,T])$ such that $q(T) = q_t(T) = 0, \quad q(0) = 1,$ and $q_t(0) = 0$. It is obvious that

$$\int_0^T \frac{d^2K}{dt^2}(t)q(t)\,dt = \int_0^T L(t)q(t)\,dt, \tag{1.18}$$

and thus,

$$-\frac{dK}{dt}(0) + \int_0^T K(t)\frac{d^2q}{dt^2}(t)\,dt = \int_0^T L(t)q(t)\,dt. \tag{1.19}$$

By virtue of (1.11), it holds that

$$\begin{aligned} 0 = & -\langle u_1, \theta(x,y,z)\rangle_1 + \int_0^\infty \int_0^T \langle u,\, q_{tt}\theta - q\Delta\theta\rangle_3\,dt\,dz \\ & = -\langle u_1, \theta(x,y,z)\rangle_1 + \int_0^T K(t)q_{tt}(t)\,dt - \int_0^T L(t)q(t)\,dt. \end{aligned} \tag{1.20}$$

By combining (1.19) and (1.20), we have

$$\langle u_1,\, \theta(x,y,z)\rangle_1 \,=\, \frac{dK}{dt}(0) \,=\, \langle u_t(0,x,y,z),\, \theta(x,y,z)\rangle_1. \tag{1.21}$$

By choosing a different $q(t)$, it also holds that

$$\langle u_0,\, \theta(x,y,z)\rangle_2 \,=\, \langle u(0,x,y,z),\, \theta(x,y,z)\rangle_2. \tag{1.22}$$

Hence, u satisfies the initial conditions.

PROPOSITION 1.6. *Let $u_0 \in L^2(0,\infty\,;\, H^{-1}(R^2))$ and $u_1 \in \mathcal{X}_1$. Then, there is a unique solution of* (1.9) *in $C\big([0,T];\, L^2(0,\infty;\, H^{-1}(R^2))\big) \cap C^1([0,T];\, \mathcal{X}_1)$.*

Proof. We can choose a sequence of functions $\{u_0^n\}_{n=1}^\infty$ such that $u_0^n \in C_0^\infty(R^3_+)$ and

$$u_0^n \to u_0 \qquad \text{in} \quad L^2(0,\infty\,;\, H^{-1}(R^2)).$$

Since $u_1 \in \mathcal{X}_1$, we can write

$$u_1 = \frac{df}{dz} + g \tag{1.23}$$

for some $f \in L^2(0,\infty\,;\, H^{-1}(R^2))$ and $g \in L^2(0,\infty\,;\, H^{-2}(R^2))$. Let $\{f_n\}_{n=1}^\infty$ and $\{g_n\}_{n=1}^\infty$ be sequences in $C_0^\infty(R^3_+)$ such that

$$f_n \to f \qquad \text{in} \quad L^2(0,\infty\,;\, H^{-1}(R^2)),$$

$$g_n \to g \qquad \text{in} \quad L^2(0,\infty\,;\, H^{-2}(R^2))$$

as $n \to \infty$. We now define

$$v_0^n(x,y,z) = \begin{cases} u_0^n(x,y,z) & \text{for } z \geq 0, \\ -u_0^n(x,y,-z) & \text{for } z < 0; \end{cases} \tag{1.24}$$

$$\tilde{f}_n(x,y,z) = \begin{cases} f_n(x,y,z) & \text{for } z \geq 0, \\ f_n(x,y,-z) & \text{for } z < 0; \end{cases} \tag{1.25}$$

$$\tilde{g}_n(x,y,z) = \begin{cases} g_n(x,y,z) & \text{for } z \geq 0, \\ -g_n(x,y,-z) & \text{for } z < 0; \end{cases} \tag{1.26}$$

$$v_1^n(x,y,z) = \frac{d\tilde{f}_n}{dz}(x,y,z) + \tilde{g}_n(x,y,z). \tag{1.27}$$

Then, it is easy to see that as $n \to \infty$,

$$v_0^n \text{ converges in } L^2(R_z\,;\, H^{-1}(R^2)) \tag{1.28}$$

and

$$v_1^n \text{ converges in } \mathcal{X}. \tag{1.29}$$

With (u_0, u_1) replaced by (v_0^n, v_1^n) in the Cauchy problem (1.1), we obtain a solution $v^n \in C([0,T]\,;\, H^m(R^3)) \cap C^1([0,T]\,;\, H^m(R^3))$ for every $m \in R$. By Lemma 1.4, (1.28), and (1.29), we find that

$$v^n \to v \qquad \text{in} \quad C\big([0,T]\,;\, L^2(R_z; H^{-1}(R^2))\big) \cap C^1([0,T]\,;\, \mathcal{X}) \tag{1.30}$$

for some v. Since v_0^n and v_1^n are odd in z, it follows that v^n is odd in z. So $v^n(t,x,y,0) = 0$. Since v^n is a smooth solution, it is evident that the restriction of v^n to R^3_+ satisfies (1.11) with (u_0, u_1) replaced by the restriction of (v_0^n, v_1^n) to R^3_+. Now it follows from (1.30) that the restriction of v to R^3_+ is a solution of (1.9). For the uniqueness of the solution, choose an arbitrary function $\psi(t,x,y,z) \in C_0^\infty((0,T)\times R^3_+)$, and find the solution of

$$\begin{cases} \phi_{tt} - \Delta\phi = \psi & \text{in} \quad (0,T)\times R^3_+, \\ \phi(t,x,y,0) = 0 & \text{in} \quad (0,T)\times R^2, \\ \phi(T,x,y,z) = 0 & \text{in} \quad R^3_+, \\ \phi_t(T,x,y,z) = 0 & \text{in} \quad R^3_+. \end{cases} \tag{1.31}$$

Then ϕ is eligible as a test function for (1.11). Let w be the difference between two solutions satisfying (1.11). Then, it is apparent that

$$\int_0^\infty \int_0^T \langle w,\, \psi\rangle_3 \, dt\, dz = 0, \tag{1.32}$$

and hence, $w \equiv 0$. □

As mentioned earlier, some regularity conditions are necessary for the uniqueness of solution of (1.9). To see this, let us consider

$$u = \delta(t - z), \tag{1.33}$$

where δ is the Dirac delta measure. Then, u is obviously a weak solution of the wave equation and satisfies the homogeneous initial-boundary conditions in the following sense.

$$\lim_{t\to 0+} \langle u, \phi(x,y,z)\rangle = 0, \tag{1.34}$$

$$\lim_{t\to 0+} \langle u_t, \phi(x,y,z)\rangle = 0, \tag{1.35}$$

$$\lim_{z\to 0+} \langle u, \psi(t,x,y)\rangle = 0 \tag{1.36}$$

for every $\phi \in C_0^\infty(R^3_+)$ and $\psi \in C_0^\infty((0,T)\times R^2)$. The bracket denotes the duality pairing between distributions and test functions. In fact, this also satisfies the boundary condition $\mathcal{B}u = 0$ at $z = 0$ in the sense

$$\lim_{z\to 0+} \langle \mathcal{B}u,\, \psi(t,x,y)\rangle = 0 \tag{1.37}$$

for all $\psi \in C_0^\infty((0,T)\times R^2)$.

2. Proof of Theorem 0.1. We first outline the strategy of proof.

Step 1. For given $0 < T < \infty$, $u_0 \in H^2(R^3_+)$, and $u_1 \in H^1(R^3_+)$ with $\frac{\partial}{\partial z}u_0(x,y,0) \in H^1(R^2)$, we find $\tilde{u}_0 \in H^2(R^3)$ and $\tilde{u}_1 \in H^1(R^3)$ which satisfy the following conditions:

$$\tilde{u}_0 = u_0, \qquad \tilde{u}_1 = u_1 \qquad \text{for} \quad z \geq 0; \tag{2.1}$$

$$\mathcal{B}\tilde{u}_0 \in L^2(R_z\,;\, H^{-1}(R^2)), \qquad \mathcal{B}\tilde{u}_1 \in \mathcal{X}; \tag{2.2}$$

(2.3) there are sequences $\{w_0^n\}_{n=1}^\infty$, $\{w_1^n\}_{n=1}^\infty$ in $C_0^\infty(R^3)$ such that

(i) $$w_0^n \to \mathcal{B}\tilde{u}_0 \qquad \text{in} \quad L^2(R_z\,;\, H^{-1}(R^2)),$$

(ii) $$w_1^n \to \mathcal{B}\tilde{u}_1 \qquad \text{in} \quad \mathcal{X},$$

(iii)
$$w_0^n(x,y,z) = -w_0^n(x,y,-z), \quad w_1^n(x,y,z) = -w_1^n(x,y,-z) \quad \text{for } -2T < z < 2T.$$

Step 2. We proceed as in Proposition 1.6 with help of Lemmas 1.1 and 1.2. Find the solution w^n of the Cauchy problem (1.1) in $(0,2T)\times R^3$ with the initial data (w_0^n, w_1^n). Let w be the limit of $\{w^n\}_{n=1}^\infty$. Then, the restriction of w to $[0,T]\times R^3_+$ is a solution of (1.9) with initial data $(\mathcal{B}u_0, \mathcal{B}u_1)$.

Step 3. Find the solution u of the Cauchy problem (1.1) in $(0,2T)\times R^3$ with the initial data $(\tilde{u}_0, \tilde{u}_1)$. By the uniqueness of solution of (1.1), we have $w \equiv \mathcal{B}u$, and the restriction of u to $[0,T]\times R^3_+$ is a desired solution.

Step 4. Prove the uniqueness of solution of (0.1) - (0.3).

One might wonder if we can bypass Step 1 and use Proposition 1.6 directly with initial data $(\mathcal{B}u_0, \mathcal{B}u_1)$ to satisfy the boundary condition $\mathcal{B}u = 0$ at $z = 0$. This is not possible since the solution obtained by Proposition 1.6 cannot be properly related to the solution required in Theorem 0.1. The trouble lies in inverting the operator $\mathcal{B}$. We now present the details of the proof. Obviously, only Step 1 and Step 4 require technical details.

Construction of $\tilde{u}_0$ and $\tilde{u}_1$. Suppose that $u_0 \in H^2(R^3_+)$ is given with $\frac{\partial}{\partial z}u_0(x,y,0) \in H^1(R^2)$. Let us consider the following initial value problem:

$$\frac{\partial^2\hat{\phi}}{\partial s^2} + |\xi|^2\frac{\partial\hat{\phi}}{\partial s} - |\xi|^2\hat{\phi} = 2|\xi|^2\frac{\partial\hat{u}_0}{\partial s}(\xi,s), \qquad s>0, \tag{2.4}$$

$$\hat{\phi}(\xi,0) = 2\hat{u}_0(\xi,0), \tag{2.5}$$

$$\frac{\partial\hat{\phi}}{\partial s}(\xi,0) = 0, \tag{2.6}$$

where $\xi=(\xi_1,\xi_2)$ is the dual variable of (x,y) and $\hat{u}_0$ denotes the Fourier transform of u_0 with respect to (x,y). Here ξ is a parameter in this initial value problem, which is well posed for almost all $\xi \in R^2$. It is easy to find a solution $\hat{\phi}$ of (2.4)–(2.6) by the variation of constants formula:

$$\begin{aligned}\hat{\phi}(\xi,s) &= 2\int_0^s \frac{e^{\lambda_1(s-\eta)}}{\sqrt{|\xi|^4+4|\xi|^2}}|\xi|^2\frac{\partial\hat{u}_0}{\partial\eta}(\xi,\eta)\,d\eta \\ &\quad - 2\int_0^s \frac{e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4+4|\xi|^2}}|\xi|^2\frac{\partial\hat{u}_0}{\partial\eta}(\xi,\eta)\,d\eta \\ &\quad - \frac{2(\lambda_2/\lambda_1)e^{\lambda_1 s}}{1-(\lambda_2/\lambda_1)}\hat{u}_0(\xi,0) \\ &\quad + \frac{2e^{\lambda_2 s}}{1-(\lambda_2/\lambda_1)}\hat{u}_0(\xi,0) \\ &= \hat{J}_1(\xi,s)+\hat{J}_2(\xi,s)+\hat{J}_3(\xi,s)+\hat{J}_4(\xi,s),\end{aligned} \tag{2.7}$$

where

$$\begin{cases}\lambda_1 = \dfrac{1}{2}(-|\xi|^2+\sqrt{|\xi|^4+4|\xi|^2}), \\ \lambda_2 = \dfrac{1}{2}(-|\xi|^2-\sqrt{|\xi|^4+4|\xi|^2}),\end{cases} \tag{2.8}$$

and

$$\hat{J}_1(\xi,s) = 2\int_0^s \frac{\lambda_1 e^{\lambda_1(s-\eta)}}{\sqrt{|\xi|^4+4|\xi|^2}}|\xi|^2\hat{u}_0(\xi,\eta)\,d\eta, \tag{2.9}$$

$$\hat{J}_2(\xi,s) = \frac{2|\xi|^2}{\sqrt{|\xi|^4+4|\xi|^2}}\hat{u}_0(\xi,s) + \frac{2\lambda_1 e^{\lambda_1 s}}{\sqrt{|\xi|^4+4|\xi|^2}}\hat{u}_0(\xi,0), \tag{2.10}$$

$$\hat{J}_3(\xi,s) = -2\int_0^s \frac{e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4+4|\xi|^2}}|\xi|^2\frac{\partial\hat{u}_0}{\partial\eta}(\xi,\eta)\,d\eta, \tag{2.11}$$

$$\hat{J}_4(\xi,s) = \frac{2e^{\lambda_2 s}}{1-(\lambda_2/\lambda_1)}\hat{u}_0(\xi,0). \tag{2.12}$$

We fix a function $\rho \in C_0^\infty(R)$ such that

$$\rho(s) = 1 \qquad \text{for } |s| \le 2T \tag{2.13}$$

and estimate each $\rho(s)\hat{J}_i(\xi, s)$. There is a positive constant M such that for all $\xi \in R^2$,

$$0 \le \lambda_1 \le M, \tag{2.14}$$

$$|\lambda_2 + |\xi|^2| \le M. \tag{2.15}$$

By means of (2.14) and the inequality

$$\|(1+|\xi|)^{3/2}\hat{u}_0(\xi,0)\|_{L^2(R^2)} \le M\|u_0\|_{H^2(R^3_+)}, \tag{2.16}$$

it is evident that for $i = 1, 2$,

$$\int_0^\infty \left(\left\| \frac{\partial^2}{\partial s^2}(\rho(s)\hat{J}_i(\xi,s)) \right\|^2_{L^2(R^2)} + \left\| \rho(s)(1+|\xi|^2)\hat{J}_i(\xi,s) \right\|^2_{L^2(R^2)} \right) ds \le M\|u_0\|^2_{H^2(R^3_+)} \tag{2.17}$$

for some positive constant M. Next we recall some basic estimates of solutions of the heat equation, which follow directly from energy estimates:

$$\int_0^\infty \left\| e^{-|\xi|^2 s}\, |\xi|^2\, \hat{f}(\xi) \right\|^2_{L^2(R^2)} ds \le M\|f\|^2_{H^1(R^2)} \quad \text{for all } f \in H^1(R^2) \tag{2.18}$$

and

$$\int_0^\infty \left\| \int_0^s e^{-|\xi|^2(s-t)}\, |\xi|^2\, \hat{g}(\xi,t)\, dt \right\|^2_{L^2(R^2)} ds \le M\|g\|^2_{L^2(R^3_+)} \quad \text{for all } g \in L^2(R^3_+) \tag{2.19}$$

for some positive constant M. We can write $\lambda_2 = -|\xi|^2 + p(\xi)$ and

$$\hat{J}_4(\xi, s) = \frac{2e^{-|\xi|^2 s}}{1 - (\lambda_2/\lambda_1)}\, e^{p(\xi)s}\, \hat{u}_0(\xi, 0), \tag{2.20}$$

where $p(\xi)$ is uniformly bounded, which follows from (2.15). By virtue of (2.14), (2.16), and (2.18), it is apparent that

$$\int_0^\infty \left(\left\| \frac{\partial^2}{\partial s^2}(\rho(s)\hat{J}_4(\xi,s)) \right\|^2_{L^2(R^2)} + \left\| (1+|\xi|^2)\rho(s)\hat{J}_4(\xi,s) \right\|^2_{L^2(R^2)} \right) ds \le M\|u_0\|^2_{H^2(R^3_+)}. \tag{2.21}$$

We rewrite $\dfrac{\partial}{\partial s}\hat{J}_3$ as

$$\begin{aligned} \frac{\partial}{\partial s}\hat{J}_3(\xi,s) = & -2\frac{|\xi|^2 e^{\lambda_2 s}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \frac{\partial \hat{u}_0}{\partial s}(\xi, 0) \\ & -2\int_0^s \frac{|\xi|^2 e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \frac{\partial^2 \hat{u}_0}{\partial \eta^2}(\xi,\eta)\, d\eta, \\ & = \hat{I}_1(\xi,s) + \hat{I}_2(\xi,s), \end{aligned} \tag{2.22}$$

where $\hat{I}_1$ is the first term of the right-hand side, and $\hat{I}_2$ is the integral term. By the same argument as for $\hat{J}_4$, we have

$$\int_0^\infty \left\| \frac{\partial(\rho \hat{I}_1)}{\partial s}(\xi, s) \right\|^2_{L^2(R^2)} ds \le M \| \frac{\partial u_0}{\partial z}(x, y, 0) \|^2_{H^1(R^2)} \tag{2.23}$$

for some positive constant M. $\hat{I}_2$ can be written as

$$\hat{I}_2(\xi, s) = -2e^{p(\xi)s} \int_0^s \frac{|\xi|^2 e^{-|\xi|^2(s-\eta)}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \, e^{-p(\xi)\eta} \, \frac{\partial^2 \hat{u}_0}{\partial \eta^2}(\xi, \eta) \, d\eta. \tag{2.24}$$

Since $p(\xi)$ is uniformly bounded, we can derive by (2.19)

$$\int_0^\infty \left\| \frac{\partial}{\partial s}(\rho(s) \hat{I}_2(\xi, s)) \right\|^2_{L^2(R^2)} ds \le M \| \frac{\partial^2 u_0}{\partial z^2} \|^2_{L^2(R^3_+)} \tag{2.25}$$

for some positive constant M. By a similar argument, we can also estimate $\rho(s)(1 + |\xi|^2)\hat{J}_3(\xi, s)$ and arrive at

$$\begin{aligned} \int_0^\infty \left(\left\| \frac{\partial^2}{\partial s^2}(\rho(s) \ \hat{J}_3(\xi, s)) \right\|^2_{L^2(R^2)} + \| (1 + |\xi|^2)\rho(s)\hat{J}_3(\xi, s) \|^2_{L^2(R^2)} \right) ds \\ \le M \left(\|u_0\|^2_{H^2(R^3_+)} + \left\| \frac{\partial u_0}{\partial z}(x, y, 0) \right\|^2_{H^1(R^2)} \right). \end{aligned} \tag{2.26}$$

Combining (2.17), (2.21), and (2.26), we have

$$\begin{aligned} \int_0^\infty \left(\left\| \frac{\partial^2}{\partial s^2}(\rho(s) \ \hat{\phi}(\xi, s)) \right\|^2_{L^2(R^2)} + \| (1 + |\xi|^2)\rho(s)\hat{\phi}(\xi, s) \|^2_{L^2(R^2)} \right) ds \\ \le M \left(\|u_0\|^2_{H^2(R^3_+)} + \left\| \frac{\partial u_0}{\partial z}(x, y, 0) \right\|^2_{H^1(R^2)} \right), \end{aligned} \tag{2.27}$$

where M is a positive constant depending on T through (2.13). We now define $\tilde{u}_0$ by

$$\tilde{u}_0(x, y, z) = \begin{cases} u_0(x, y, z) & \text{for } z \ge 0, \\ \rho(z)\phi(x, y, -z) - u_0(x, y, -z) & \text{for } z < 0, \end{cases} \tag{2.28}$$

where $\phi(x, y, s)$ is the Fourier inverse transform of $\hat{\phi}(\xi, s)$ above. Then, it follows that

$$\begin{cases} \lim\limits_{z \to 0+} \tilde{u}_0(\cdot, z) = \lim\limits_{z \to 0-} \tilde{u}_0(\cdot, z), \\ \lim\limits_{z \to 0+} \dfrac{\partial}{\partial z} \tilde{u}_0(\cdot, z) = \lim\limits_{z \to 0-} \dfrac{\partial}{\partial z} \tilde{u}_0(\cdot, z), \end{cases} \tag{2.29}$$

and consequently,

$$\|\tilde{u}_0\|_{H^2(R^3)} \le M \left(\|u_0\|^2_{H^2(R^3_+)} + \left\| \frac{\partial u_0}{\partial z}(x, y, 0) \right\|^2_{H^1(R^2)} \right), \tag{2.30}$$

and

$$\mathcal{B}\tilde{u}_0 \in L^2(R_z \, ; \, H^{-1}(R^2)). \tag{2.31}$$

It also follows that, for almost all $-2T < z < 2T$,

$$(\mathcal{B}\tilde{u}_0)(\cdot,-z) = -(\mathcal{B}\tilde{u}_0)(\cdot,z) \qquad \text{in } H^{-1}(R^2). \tag{2.32}$$

Since $\mathcal{B}\tilde{u}_0 \in L^2(R_z\,;H^{-1}(R^2))$, there are sequences $\{\zeta_1^n\}_{n=1}^\infty$ in $C_0^\infty((0,2T)\times R^2)$, $\{\zeta_2^n\}_{n=1}^\infty$ in $C_0^\infty((2T,\infty)\times R^2)$, and $\{\zeta_3^n\}_{n=1}^\infty$ in $C_0^\infty((-\infty,-2T)\times R^2)$ such that as $n\to\infty$,

$$\zeta_1^n \to \chi_1\,\mathcal{B}\tilde{u}_0 \qquad \text{in } L^2(R_z\,;H^{-1}(R^2)), \tag{2.33}$$

$$\zeta_2^n \to \chi_2\,\mathcal{B}\tilde{u}_0 \qquad \text{in } L^2(R_z\,;H^{-1}(R^2)), \tag{2.34}$$

and

$$\zeta_3^n \to \chi_3\,\mathcal{B}\tilde{u}_0 \qquad \text{in } L^2(R_z\,;H^{-1}(R^2)), \tag{2.35}$$

where $\chi_1(z)$ is a characteristic function for the interval $[0,2T]$, $\chi_2(z)$ is for the interval $[2T,\infty)$, and $\chi_3(z)$ is for $(-\infty,-2T]$. We then set

$$w_0^n(x,y,z) = \zeta_1^n(x,y,z) - \zeta_1^n(x,y,-z) + \zeta_2^n(x,y,z) + \zeta_3^n(x,y,z). \tag{2.36}$$

Then, all the conditions in (2.1)–(2.3) for u_0 are satisfied.

Next we suppose that $u_1 \in H^1(R^3_+)$ is given. We consider (2.4)–(2.6) with $\hat{u}_0$ replaced by $\hat{u}_1$. Then, the solution $\hat{\psi}$ of (2.4)–(2.6) is given by

$$\hat{\psi}(\xi,s) = \hat{H}_1(\xi,s) + \hat{H}_2(\xi,s) + \hat{H}_3(\xi,s) + \hat{H}_4(\xi,s), \tag{2.37}$$

where

$$\hat{H}_1(\xi,s) = 2\int_0^s \frac{\lambda_1 e^{\lambda_1(s-\eta)}}{\sqrt{|\xi|^4+4|\xi|^2}}|\xi|^2\hat{u}_1(\xi,\eta)\,d\eta, \tag{2.38}$$

$$\hat{H}_2(\xi,s) = \frac{2|\xi|^2}{\sqrt{|\xi|^4+4|\xi|^2}}\hat{u}_1(\xi,s) + \frac{2\lambda_1 e^{\lambda_1 s}}{\sqrt{|\xi|^4+4|\xi|^2}}\hat{u}_1(\xi,0), \tag{2.39}$$

$$\hat{H}_3(\xi,s) = -2\int_0^s \frac{e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4+4|\xi|^2}}|\xi|^2\frac{\partial\hat{u}_1}{\partial\eta}(\xi,\eta)\,d\eta, \tag{2.40}$$

$$\hat{H}_4(\xi,s) = \frac{2e^{\lambda_2 s}}{1-(\lambda_2/\lambda_1)}\hat{u}_1(\xi,0). \tag{2.41}$$

Since $u_1 \in H^1(R^3_+)$ implies $u_1(x,y,0) \in H^{1/2}(R^2)$, we can easily derive

$$\int_0^\infty \left(\left\|\rho(s)\frac{\partial\hat{H}_i}{\partial s}(\xi,s)\right\|^2_{L^2(R^2)} + \left\|\rho(s)(1+|\xi|)\hat{H}_i(\xi,s)\right\|^2_{L^2(R^2)}\right) ds \le M\|u_1\|^2_{H^1(R^3_+)}, \tag{2.42}$$

for $i = 1, 2, 4$, where ρ is the same as defined by (2.13). For the estimate of $\hat{H}_3$, we write

$$\frac{\partial \hat{H}_3}{\partial s}(\xi, s) = -2\frac{|\xi|^2}{\sqrt{|\xi|^4 + 4|\xi|^2}}\frac{\partial \hat{u}_1}{\partial s}(\xi, s) - 2\int_0^s \frac{\lambda_2 |\xi|^2 e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4 + 4|\xi|^2}}\frac{\partial \hat{u}_1}{\partial \eta}(\xi, \eta)\, d\eta, \tag{2.43}$$

and, by means of (2.15) and (2.19), we find that

$$\int_0^\infty \left\| \rho(s)\frac{\partial \hat{H}_3}{\partial s}(\xi, s) \right\|^2_{L^2(R^2)} ds \le M\|u_1\|^2_{H^1(R^3_+)} \tag{2.44}$$

for some positive constant M. It also holds that

$$\int_0^\infty \left\| \rho(s)(1 + |\xi|)\hat{H}_3(\xi, s) \right\|^2_{L^2(R^2)} ds \le M\|u_1\|^2_{H^1(R^3_+)}. \tag{2.45}$$

Let us define

$$\tilde{u}_1(x, y, z) = \begin{cases} u_1(x, y, z) & \text{for } z \ge 0, \\ \rho(z)\psi(x, y, -z) - u_1(x, y, -z) & \text{for } z < 0. \end{cases} \tag{2.46}$$

Then we have

$$\lim_{z \to 0+} \tilde{u}_1(\cdot, z) = \lim_{z \to 0-} \tilde{u}_1(\cdot, z) \quad \text{in } H^{1/2}(R^2), \tag{2.47}$$

which, together with (2.42), (2.44), and (2.45), yields

$$\frac{\partial \tilde{u}_1}{\partial z} \in L^2(R^3) \tag{2.48}$$

and

$$\|\tilde{u}_1\|_{H^1(R^3)} \le M\|u_1\|_{H^1(R^3_+)}, \tag{2.49}$$

for some positive constant M depending on T through (2.13). Next we choose a sequence $\{v^m\}_{m=1}^\infty$ in $H^3(R^3_+)$ such that

$$v^m \to u_1 \quad \text{in } H^1(R^3_+) \tag{2.50}$$

as $m \to \infty$. Let $\hat{\psi}^m$ be the solution of (2.4)–(2.6) with $\hat{u}_0$ replaced by $\hat{v}^m$, and define

$$\tilde{v}^m(x, y, z) = \begin{cases} v^m(x, y, z) & \text{for } z \ge 0, \\ \rho(z)\psi^m(x, y, -z) - v^m(x, y, -z) & \text{for } z < 0. \end{cases} \tag{2.51}$$

Then, by virtue of (2.49), we find that

$$\mathcal{B}\tilde{v}^m \to \mathcal{B}\tilde{u}_1 \quad \text{in } \mathcal{X}. \tag{2.52}$$

In the meantime, the result for u_0 implies that each $\mathcal{B}\tilde{v}^m$ can be approximated in $L^2(R_z\,;\,H^{-1}(R^2))$ by a sequence of functions in $C_0^\infty(R^3)$ which are odd in $z \in (-2T,\,2T)$. Hence, there is a sequence $\{w_1^n\}_{n=1}^\infty$ in $C_0^\infty(R^3)$ such that

$$w_1^n \to \mathcal{B}\tilde{u}_1 \qquad \text{in } \mathcal{X}, \tag{2.53}$$

and

$$w_1^n(x,y,-z) = -w_1^n(x,y,-z) \qquad \text{for } -2T < z < 2T. \tag{2.54}$$

Obviously, all the conditions (2.1)-(2.3) for u_1 are satisfied.

Solutions of (1.1) *with initial data* $(\tilde{u}_0, \tilde{u}_1)$. Consider the Cauchy problem (1.1) with the initial data (w_0^n, w_1^n). For each n, there is a unique solution $w^n \in C([0,2T]\,;\,H^m(R^3)) \cap C^1([0,2T]\,;\,H^{m-1}(R^3))$ for every $m \in R$. By Lemma 1.2 and the condition (iii) in (2.3), $w^n(t,x,y,0) = 0$ for all $(t,x,y) \in [0,T]\times R^2$. Hence, the restriction of w^n to $[0,T]\times R^3_+$ satisfies (1.11) with (u_0,u_1) replaced by the restriction of (w_0^n, w_1^n) to $z>0$. Meanwhile, we have, by Lemma 1.4,

$$w^n \to \mathcal{B}\tilde{u} \qquad \text{in } C\big([0,T]\,;\,L^2(R_z; H^{-1}(R^2))\big) \cap C^1([0,T]\,;\,\mathcal{X}), \tag{2.55}$$

where $\tilde{u}$ is the unique solution of (1.1) with the initial data $(\tilde{u}_0, \tilde{u}_1)$. Hence, the restriction of $\mathcal{B}\tilde{u}$ to $[0,T]\times R^3_+$ satisfies (1.11) with (u_0,u_1) replaced by the restriction of $(\mathcal{B}u_0, \mathcal{B}u_1)$ to $z>0$. Consequently, the restriction of $\tilde{u}$ to $[0,T]\times R^3_+$ is a desired solution in Theorem 0.1.

Uniqueness of the solution. Let Φ be the difference between two solutions in Theorem 0.1. Then, we have

$$\Phi \in C([0,T]\,;\,H^2(R^3_+)) \cap C^1([0,T]\,;\,H^1(R^3_+)), \tag{2.56}$$

$$\Phi(0,x,y,z) = 0, \qquad \Phi_t(0,x,y,z) = 0 \qquad \text{for } (x,y,z) \in R^3_+, \tag{2.57}$$

$$\Phi_{tt} - \Delta\Phi = 0 \qquad \text{in } (0,T)\times R^3_+. \tag{2.58}$$

Since $\mathcal{B}\Phi$ is a solution of (1.9) with

$$(\mathcal{B}\Phi)(0,\cdot) = 0, \qquad (\mathcal{B}\Phi_t)(0,\cdot) = 0, \tag{2.59}$$

it follows from Proposition 1.6 that

$$\mathcal{B}\Phi \equiv 0 \qquad \text{in } (0,T)\times R^3_+. \tag{2.60}$$

By virtue of (2.57) and the domain of dependence of the solution, we also have

$$\Phi = 0 \qquad \text{for } 0 \le t < z. \tag{2.61}$$

Since $\Phi \in C([0,T]\,;\,H^2(R^3_+))$, (2.60) implies that for each $t \in [0,T]$, $\xi \in R^2$,

$$\frac{\partial^2\hat{\Phi}}{\partial z^2}(t,\xi,z) - |\xi|^2\frac{\partial\hat{\Phi}}{\partial z}(t,\xi,z) - |\xi|^2\hat{\Phi}(t,\xi,z) = 0 \tag{2.62}$$

holds in $\mathcal{D}'((0,\infty))$. But it follows from (2.61) that

$$\hat{\Phi}(t,\xi,z) = 0 \qquad \text{for all } \xi \in R^2,\ 0 \le t < z, \tag{2.63}$$

which, together with (2.62), yields

$$\hat{\Phi} \equiv 0 \qquad \text{in } R^3_+. \tag{2.64}$$

Hence, the solution of Theorem 0.1 is unique.

Finally, the estimate (0.10) follows from Lemma 1.1, (2.30), and (2.49).

3. Proof of corollaries. For the proof of corollaries, it is enough to obtain necessary estimates of $\hat{\phi}(\xi, s)$ given by (2.7). We start from Corollary 0.2. Suppose $u_0 \in H^3(R^3_+)$ and $\mathcal{B}u_0 = 0$ at $z = 0$. By the same argument used before, it is easy to see for $i = 1, 2$,

(3.1)
$$\int_0^\infty \left(\left\| \frac{\partial^3}{\partial s^3}(\rho(s)\hat{J}_i(\xi, s)) \right\|^2_{L^2(R^2)} + \left\| \rho(s)(1+|\xi|)^3 \hat{J}_i(\xi, s) \right\|^2_{L^2(R^2)} \right) ds \le M\|u_0\|^2_{H^3(R^3_+)},$$

where $\rho(s)$ is the same as before. By means of (2.12) and (2.22), we write

$$\begin{aligned} \frac{\partial^2 \hat{J}_3}{\partial s^2}(\xi, s) + \frac{\partial^2 \hat{J}_4}{\partial s^2}(\xi, s) &= \frac{2\lambda_1\lambda_2^2 e^{\lambda_2 s}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \hat{u}_0(\xi, 0) \\ &\quad - 2\frac{\lambda_2|\xi|^2 e^{\lambda_2 s}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \frac{\partial \hat{u}_0}{\partial s}(\xi, 0) - 2\frac{|\xi|^2 e^{\lambda_2 s}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \frac{\partial^2 \hat{u}_0}{\partial s^2}(\xi, 0) \\ &\quad - 2\int_0^s \frac{|\xi|^2 e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \frac{\partial^3 \hat{u}_0}{\partial \eta^3}(\xi, \eta)\, d\eta. \end{aligned} \tag{3.2}$$

However, the first three terms in the right-hand side can be written as

$$\begin{aligned} -2\frac{|\xi|^2 e^{\lambda_2 s}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \left(\frac{\partial^2 \hat{u}_0}{\partial s^2}(\xi, 0) - |\xi|^2 \frac{\partial \hat{u}_0}{\partial s}(\xi, 0) - |\xi|^2 \hat{u}_0(\xi, 0) \right) \\ + e^{\lambda_2 s} \left(p_1(\xi) \frac{\partial \hat{u}_0}{\partial s}(\xi, 0) + p_2(\xi)\hat{u}_0(\xi, 0) \right), \end{aligned} \tag{3.3}$$

where $p_1(\xi)$ and $p_2(\xi)$ are uniformly bounded in ξ. Since $\mathcal{B}u_0 = 0$ at $z = 0$, (3.2) is reduced to

$$\begin{aligned} \frac{\partial^2 \hat{J}_3}{\partial s^2}(\xi, s) + \frac{\partial^2 \hat{J}_4}{\partial s^2}(\xi, s) &= e^{\lambda_2 s} \left(p_1(\xi) \frac{\partial \hat{u}_0}{\partial s}(\xi, 0) + p_2(\xi)\hat{u}_0(\xi, 0) \right) \\ &\quad - 2\int_0^s \frac{|\xi|^2 e^{\lambda_2(s-\eta)}}{\sqrt{|\xi|^4 + 4|\xi|^2}} \frac{\partial^3 \hat{u}_0}{\partial \eta^3}(\xi, \eta)\, d\eta. \end{aligned} \tag{3.4}$$

It now follows from (2.18) and (2.19) that

$$\int_0^\infty \left\| \rho(s) \left(\frac{\partial^3 \hat{J}_3}{\partial s^3}(\xi, s) + \frac{\partial^3 \hat{J}_4}{\partial s^3}(\xi, s) \right) \right\|^2_{L^2(R^2)} ds \le M\|u_0\|^2_{H^3(R^3_+)}. \tag{3.5}$$

We can easily find that

$$\int_0^\infty \left\| \rho(s)(1+|\xi|)^3 (\hat{J}_3(\xi, s) + \hat{J}_4(\xi, s)) \right\|^2_{L^2(R^2)} ds \le M\|u_0\|^2_{H^3(R^3_+)}. \tag{3.6}$$

By virtue of (3.1) and (3.6), we can estimate ϕ;

$$\|\rho(z)\phi(x, y, z)\|_{H^3(R^3_+)} \le M\|u_0\|_{H^3(R^3_+)}, \tag{3.7}$$

for some positive constant M, which depends on T through (2.13). Recalling (2.4), (2.28), (2.29), and the boundary condition $\mathcal{B}u_0 = 0$ at $z = 0$ we derive

$$\lim_{z \to 0+} \frac{\partial^2 \tilde{u}_0}{\partial z^2}(\cdot, z) = \lim_{z \to 0-} \frac{\partial^2 \tilde{u}_0}{\partial z^2}(\cdot, z), \tag{3.8}$$

which, combined with (3.7), yields

$$\|\tilde{u}_0\|_{H^3(R^3)} \le M\|u_0\|_{H^3(R^3_+)}. \tag{3.9}$$

The estimate of $\tilde{u}_1$ defined by (2.46) follows directly from the estimate of $\tilde{u}_0$ in the previous section. This completes the proof of Corollary 0.2.

For the proof of the remaining corollaries, we first observe the following fact.

LEMMA 3.1. *Let $\hat{\phi}$ be a solution of* (2.4)–(2.6). *Assume that $u_0 \in H^{2m+4}(R^3_+)$, and $\mathcal{B}u_0 = 0$, $\Delta\mathcal{B}u_0 = 0, \dots, \Delta^m\mathcal{B}u_0 = 0$ at $z = 0$, where m is a nonnegative integer. Then, we have*

$$\frac{\partial^{2m+2}\hat{\phi}}{\partial s^{2m+2}}(\xi,0) = 2\,\frac{\partial^{2m+2}\hat{u}_0}{\partial s^{2m+2}}(\xi,0) \quad \textit{for almost all } \xi \in R^2, \tag{3.10}$$

$$\frac{\partial^{2m+3}\hat{\phi}}{\partial s^{2m+3}}(\xi,0) = 0 \quad \textit{for almost all } \xi \in R^2. \tag{3.11}$$

Proof. Suppose that $\mathcal{B}u_0 = 0$ at $z = 0$. Then, it follows from (2.4)–(2.6) that

$$\begin{aligned}\frac{\partial^2\hat{\phi}}{\partial s^2}(\xi,0) &= 2|\xi|^2\left(\hat{u}_0(\xi,0) + \frac{\partial\hat{u}_0}{\partial s}(\xi,0)\right)\\ &= 2\frac{\partial^2\hat{u}_0}{\partial s^2}(\xi,0) \quad \text{for almost all } \xi \in R^2.\end{aligned} \tag{3.12}$$

By differentiation of (2.4) in s, and using (3.12), we obtain

$$\frac{\partial^3\hat{\phi}}{\partial s^3}(\xi,0) = 0 \qquad \text{for almost all } \xi \in R^2. \tag{3.13}$$

Suppose that Lemma 3.1 is true for $0 \le m \le k$. By differentiation of (2.4), we have

$$\frac{\partial^{2k+4}\hat{\phi}}{\partial s^{2k+4}} + |\xi|^2\frac{\partial^{2k+3}\hat{\phi}}{\partial s^{2k+3}} - |\xi|^2\frac{\partial^{2k+2}\hat{\phi}}{\partial s^{2k+2}} = 2|\xi|^2\frac{\partial^{2k+3}\hat{u}_0}{\partial s^{2k+3}}, \tag{3.14}$$

which, combined with (3.10) and (3.11) for $m = k$, yields

$$\begin{aligned}\frac{\partial^{2k+4}\hat{\phi}}{\partial s^{2k+4}}(\xi,0) &= 2|\xi|^2\left(\frac{\partial^{2k+3}\hat{u}_0}{\partial s^{2k+3}}(\xi,0) + \frac{\partial^{2k+2}\hat{u}_0}{\partial s^{2k+2}}(\xi,0)\right)\\ &= 2\,\frac{\partial^{2k+4}\hat{u}_0}{\partial s^{2k+4}}(\xi,0) \quad \text{for almost all } \xi \in R^2.\end{aligned} \tag{3.15}$$

Here we also used the fact that $\Delta^m\mathcal{B}u_0 = 0$, $0 \le m \le k+1$ at $z = 0$ implies $\frac{\partial^{2m}}{\partial z^{2m}}\mathcal{B}u_0 = 0$, $0 \le m \le k+1$, at $z = 0$. Now (3.10) is true for $m = k+1$. By differentiation of (3.14), we also get (3.11) for $m = k+1$. The proof is complete. □

Under the same assumption as in Lemma 3.1, we consider the initial value problem for $0 \le k \le m+1$,

$$\frac{\partial^2\hat{\psi}_k}{\partial s^2} + |\xi|^2\frac{\partial\hat{\psi}_k}{\partial s} - |\xi|^2\hat{\psi}_k = 2|\xi|^2\frac{\partial^{2k+1}\hat{u}_0}{\partial s^{2k+1}}(\xi,s), \qquad s > 0, \tag{3.16}$$

$$\hat{\psi}_k(\xi,0) = 2\frac{\partial^{2k}\hat{u}_0}{\partial s^{2k}}(\xi,0), \tag{3.17}$$

$$\frac{\partial\hat{\psi}_k}{\partial s}(\xi,0) = 0. \tag{3.18}$$

By the above lemma, and the uniqueness of solution of (2.4)–(2.6), we find

$$\hat{\psi}_k \equiv \frac{\partial^{2k}\hat{\phi}}{\partial s^{2k}} \qquad \text{for } 0 \le k \le m+1. \tag{3.19}$$

Recalling (2.28), we derive

$$\lim_{z\to 0+}\frac{\partial^\nu}{\partial z^\nu}\tilde{u}_0(\cdot,z) = \lim_{z\to 0-}\frac{\partial^\nu}{\partial z^\nu}\tilde{u}_0(\cdot,z), \quad 0 \le \nu \le 2m+3. \tag{3.20}$$

Meanwhile, (2.27) yields

$$\begin{aligned}\int_0^\infty \left(\left\|\frac{\partial^2}{\partial s^2}(\rho(s)\ \hat{\psi}_{m+1}(\xi,s))\right\|^2_{L^2(R^2)} + \left\|(1+|\xi|^2)\rho(s)\hat{\psi}_{m+1}(\xi,s)\right\|^2_{L^2(R^2)}\right) ds \\ \le M\left(\|u_0\|^2_{H^{2m+4}(R^3_+)} + \left\|\frac{\partial^{2m+3}\hat{u}_0}{\partial z^{2m+3}}(x,y,0)\right\|^2_{H^1(R^2)}\right).\end{aligned} \tag{3.21}$$

By means of (2.18) and (2.19), we can obtain directly from (2.9)–(2.12)

$$\int_0^\infty \|\rho(s)(1+|\xi|)^{2m+4}\hat{J}_i(\xi,s)\|^2_{L^2(R^2)}\,ds \le M\|u_0\|^2_{H^{2m+4}(R^3_+)}, \qquad i = 1,\dots,4. \tag{3.22}$$

By (3.20)–(3.22), we arrive at

$$\|\tilde{u}_0\|_{H^{2m+4}(R^3)} \le M\left(\|u_0\|_{H^{2m+4}(R^3_+)} + \left\|\frac{\partial^{2m+3}u_0}{\partial z^{2m+3}}(x,y,0)\right\|_{H^1(R^2)}\right). \tag{3.23}$$

When $u_0 \in H^{2m+3}(R^3_+)$, with $\mathcal{B}u_0 = 0,\dots,\Delta^m\mathcal{B}u_0 = 0$ at $z = 0$, we apply the estimate (3.7) to $\psi_m = \frac{\partial^{2m}\phi}{\partial z^{2m}}$ so that

$$\|\tilde{u}_0\|_{H^{2m+3}(R^3)} \le M\|u_0\|_{H^{2m+3}(R^3_+)}. \tag{3.24}$$

Here, M is a positive constant depending on T and m. Since the esimates of $\tilde{u}_1$ are identical with those of $\tilde{u}_0$, the proof of Corollaries 0.3 and 0.4 is complete. Corollary 0.5 follows trivially from the previous ones.

Acknowledgment. I would like to thank Shuming Sun for a helpful discussion.

REFERENCES

[1] J. Bergh and J. Löfström, *Interpolation Spaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[2] H. Bondi, *Waves on the surface of a compressible liquid*, Proc. Cambridge Phil. Soc., 43 (1947), pp. 75–95.

[3] G.F.D. Duff, *Hyperbolic differential equations and waves*, in Boundary Value Problems for Linear Evolution Partial Differential Equations, Proc. NATO Advanced Study Institute, H.G. Garnir, ed., D. Reidel, Dordrecht, Boston, 1976, pp. 27–155.
[4] L. Hörmander, *Linear Partial Differential Operators*, 4th ed., Springer-Verlag, Berlin, Heidelberg, New York, 1976.
[5] L.D. Landau and E.M. Lifshitz, *Fluid Mechanics*, Pergamon Press, London, Paris, Frankfurt, 1959.
[6] J.L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, Heidelberg, Berlin, 1972.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 72–94

EXISTENCE OF TRAVELING WAVES IN A BIODEGRADATION MODEL FOR ORGANIC CONTAMINANTS*

REGAN MURRAY† AND J. X. XIN†

Abstract. We study a biodegradation model for the time evolution of concentrations of contaminant, nutrient, and bacteria. The bacteria has a natural concentration which will increase when the nutrient reaches the substrate (contaminant). The growth utilizes nutrients and degrades the substrate. Eventually, such a process removes all the substrate and can be described by traveling wave solutions. The model consists of advection-reaction-diffusion equations for the substrate and nutrient concentrations and a rate equation (ODE) for the bacteria concentration. We first show the existence of approximate traveling wave solutions to an elliptically regularized system posed on a finite domain using degree theory and the elliptic maximum principle. To prove that the approximate solutions do not converge to trivial solutions, we construct comparison functions for each component and employ integral identities of the governing equations. This way, we derive a priori estimates of solutions independent of the length of the finite domain and the regularization parameter. The integral identities take advantage of the forms of coupling in the system and help us obtain optimal bounds on the traveling wave speed. We then extend the domain to the infinite line limit, remove the regularization, and construct a traveling wave solution for the original set of equations satisfying the prescribed boundary conditions at spatial infinities. The contaminant and nutrient profiles of the traveling waves are strictly monotone functions, while the biomass profile has a pulse shape.

Key words. biodegradation, organic contaminants, traveling waves

AMS subject classifications. 92C05, 92C45, 34B15, 35J25

PII. S0036141096313392

1. Introduction.

Thousands of chemical spills contaminate subsurface aquifers used for drinking water and agriculture in the United States. A promising technology for cleaning up subsurface organic contamination is in-situ bioremediation. This technique works by stimulating the bacteria already present in the soil to use the contaminant as a source of food, thereby transforming it into nontoxic components such as carbon dioxide and water. Among a variety of restoration technologies, in-situ bioremediation has been shown to be the most economical for remediating certain contaminants. Several experiments have shown that this method takes much less time and is less costly than the traditional pumping and filtering techniques. For example, one site was estimated to take 100 years to clean up using the pump and treat method; in-situ bioremediation took only 10 months [3]. Moreover, while many methods simply remove the contaminant from the site and thereby create a disposal problem, in-situ bioremediation has the potential to completely transform organic contaminants into neutral compounds by utilizing the indigenous bacteria [6].

In-situ bioremediation involves a complex combination of biological and chemical properties and fluid dynamics rendering scientific predictions difficult. Mathematical models are particularly useful in understanding the interplay between the various mechanisms in addition to making predictions, analyzing the significance of physical

*Received by the editors December 9, 1996; accepted for publication (in revised form) February 10, 1998; published electronically September 29, 1998.
http://www.siam.org/journals/sima/30-1/31339.html

†Program in Applied Mathematics, Department of Mathematics, University of Arizona, Tucson, AZ 85721 (rmurray@math.arizona.edu, xin@math.arizona.edu). The research of the first author was partially supported by a graduate teaching assistantship at the Program in Applied Mathematics, University of Arizona, and by the Center of Nonlinear Studies, Los Alamos National Laboratory. The research of the second author was supported by NSF grants DMS-9302830 and DMS-9625680.

parameters, and providing theoretical interpretations of experiments. Several models for in-situ bioremediation have been presented in the literature; see [2], [5]. In this paper, we consider the model presented in [7] and [8], for it is the simplest model to capture the fundamental aspects of degradation.

The biodegradation model is

$$R_f S_t + (vS - DS_x)_x = -R_s, \quad x \in R^1, \tag{1.1}$$

$$A_t + (vA - DA_x)_x = -\gamma R_s, \tag{1.2}$$

$$M_t = Y R_s - b(M - M_0), \tag{1.3}$$

where $R_f, b, M_0, \gamma, v, D, Y$ are all positive constants. The subscripts t and x denote time and space derivatives, while the subscripts s and f are used to distinguish between the reaction term R_s and retardation factor R_f. The variables are (S, A, M) where S is the aqueous phase concentration of the electron donor (substrate), A is the electron acceptor (nutrient concentration), and M is the biomass concentration. The constant M_0 is the neutral background biomass concentration. $R_f > 1$ is the retardation factor of the electron donor (substrate) S due to sorption. The parameter b is the cell decay coefficient for the bacteria population M. γ is a coefficient equal to the mass of A utilized by the bacteria per unit mass of S degraded. D is the diffusion coefficient and Y is the yield coefficient (mass of bacterial cells produced per mass of S degraded). The parameter v is the constant pore water velocity and M_0 is the background biomass concentration.

The biodegradation rate, R_s, is given by the Monod kinetic model

$$R_s = \frac{qMSA}{(K_S + S)(K_A + A)}, \tag{1.4}$$

where q is the maximum specific rate of substrate utilization, and K_S, K_A are the half-maximum rate concentrations of S and A. Note that the system assumes linear sorption of the electron donor (represented by the term R_f), no sorption for the electron acceptor, and a constant minimal background bacteria population M_0 (as an equilibrium between cell growth and decay). In this model, microbes are attached to the soil particles and only consume aqueous phase species.

The system (1.1)–(1.3) is one of the simplest systems in the literature; however, it contains very rich phenomena. Oya and Valocchi [8] and Valocchi [10] recently studied traveling waves in this system. They observed that "bioavailability" of the chemical constituents is an important requirement for successful biodegradation. Since the microbes are assumed to be attached to solids [5], the dissolved contaminant and nutrients must be transported to the bacteria. This is represented in the above system by the water velocity v and the sorption effect $R_f > 1$. When the nutrients reach the biomass, a biologically active zone forms and propagates. This zone supports respiration, that is, nutrient consumption, contaminant degradation, and an increase in the microbial population. [8] is the first paper to propose the idea that the biologically active zone can be mathematically modeled as traveling waves. The retreating dissolved solute concentration and the advancing nutrient concentration move together as traveling fronts, while the bacteria concentration tags along as a traveling pulse.

In this paper, we establish the existence of traveling waves in (1.1)–(1.3) under the sole condition that $R_f > 1$. The traveling wave solutions are of the form

$$S = S(x - ct),\ A = A(x - ct),\ M = M(x - ct), \tag{1.5}$$

where (S, A, M) as functions of $\xi \equiv x - ct$ satisfy the boundary conditions

$$S(-\infty) = 0, \quad S(+\infty) = S^+, \ A(-\infty) = A_-, \quad A(+\infty) = 0, \ M(\pm\infty) = M_0, \tag{1.6}$$

and A_- and S^+ are prescribed positive constants. Using (1.5), (1.1)–(1.3) are transformed into

$$DS_{\xi\xi} + (R_f c - v)S_\xi = R_s, \tag{1.7}$$

$$DA_{\xi\xi} + (c - v)A_\xi = \gamma R_s, \tag{1.8}$$

$$cM_\xi = -YR_s + b(M - M_0), \tag{1.9}$$

along with the boundary conditions

$$S(-\infty) = 0, \ S(+\infty) = S^+, \ A(-\infty) = A_-, \ A(+\infty) = 0, \ M(\pm\infty) = M_0. \tag{1.10}$$

To prove the existence of a solution to (1.7)–(1.9), we follow a framework similar to that of Berestyki, Nicolaenko, and Sheurer [1] and Xin [11], and consider the system on a bounded domain $[-d, d]$, $d > 0$. We define a fixed point map such that any solution to the fixed point problem is also a solution of (1.7)–(1.9) on the bounded domain. By determining the Leray–Schauder degree of the map (see [13] for details), we show that a solution exists to the fixed point problem. Finally, we extend the solution to the entire real line. This is a well-known technique.

It is important to note that equations (1.1)–(1.3) are substantially different than the two-equation, reaction-diffusion, combustion system studied in [1] and require new ingredients for the proof of existence. In the biodegradation model, the fronts propagate as a result of distinct advective velocities rather than reaction-diffusion. We believe traveling wave fronts in advection systems have not been much explored. In this case, although the method of [1] applies, many new steps are needed. For instance, finding good bounds on the wave speed requires a substantial amount of work on the advection and reaction terms. In addition, the two advection equations cannot be manipulated to define pointwise inequalities as in [1]. Instead, we develop *integral inequalities* based on the conserved quantity $\gamma R_f S - A$ (see Lemma 3.4).

Many other differences between the biodegradation system and [1] are noteworthy. First, while the fronts in [1] are strictly monotone, the biomass profile in (1.1)–(1.3) forms a pulse. Rather than following directly from the maximum principle, new arguments are necessary to estimate the maximum of M (see Proposition 2.1). Second, the condition to fix translation invariance, $A(0) = \theta$, is not a priori optimal as it is in [1], where θ is the ignition temperature. In this paper, the value of θ is updated three times to prevent the solutions of (1.1)–(1.3) from converging to trivial solutions in the limit $d \to \infty$ (see Proposition 3.1 and Lemmas 3.3 and 3.4). In particular, much work is required to show that S does not tend to zero as the domain extends to the real line. Finally, we develop integral identities to show that the boundary conditions remain as $d \to \infty$; see Theorem 4.1 and Lemma 5.1. Using integral identities instead of pointwise estimates to control wave speeds and large space asymptotic behavior of solutions is an efficient way of handling systems with more than two equations.

The main result of the paper is the following.

THEOREM 1.1 (existence of traveling waves). *Under the condition that the retardation factor $R_f > 1$, the system* (1.7)–(1.9) *with R_s given by* (1.4) *admits a classical traveling wave solution (S, A, M, c) of the form* (1.5) *satisfying the boundary conditions in* (1.10). *Moreover,*

$$0 < S(\xi) < S^+, \ S'(\xi) > 0, \ 0 < A(\xi) < A_-, \ A'(\xi) < 0 \quad \forall \xi \in R^1, \tag{1.11}$$

$$c = \frac{v(A_- + \gamma S^+)}{A_- + \gamma R_f S^+}, \tag{1.12}$$

$$M_0 < M \le M_0 + Y\frac{(R_f - 1)A_- S^+}{A_- + \gamma S^+} \quad \forall \xi \in R^1. \tag{1.13}$$

We see from Theorem 1.1 that the profiles of substrate and nutrient concentrations (S and A) are strictly monotone functions of ξ and that the wave speed c is independent of the parameters Y, b of the M equation. The explicit formula for the wave speed c in terms of the left and right states of solutions is reminiscent of the Rankine–Hugoniot formula for viscous shock waves in conservative equations; see [12] for such traveling waves arising in solute transport problems. The maximum norm bound of the biomass M, on the other hand, depends on the parameters and boundary conditions of the (S, A) equations as well as the yield constant Y of the M equation. Numerical simulations in [8] and those of the present authors show that M forms an asymmetric pulse with one maximum. The condition $R_f > 1$ physically means that the advective velocity of S is slower than that of A; hence, the two concentrations mix, which is essential for the biomass to grow and the three components to travel together. As pointed out in [8], there are no traveling waves if $R_f = 1$. In fact, it is obvious from (1.13) that $M \equiv M_0$ if $R_f = 1$, and no traveling pulse can form in M. It remains an interesting problem to find out if the traveling waves are unique up to a constant translation in ξ and if M always achieves one maximum.

The main result is significant, for it shows that there is a simple, well-understood solution to a very complicated phenomena. This solution occurs under a particular set of parameters; however, numerical studies suggest that this is the most common solution. In [8], the authors find waves with oscillating front shapes in time; this proof will be left for a future paper. Finally, many of the important factors required for the implementation of in-situ bioremediation are determined in Theorem 1.1, such as the speed of the traveling fronts and the maximum and minimum concentrations of the biomass.

The second result of the paper is the following.

THEOREM 1.2 (existence and uniqueness of traveling waves). *The system* (1.7)–(1.9) *with* $D = 0$ *and* R_s *given by* (1.4) *admits a* unique*, classical traveling wave solution* (S, A, M, c) *of the form* (1.5) *satisfying the boundary conditions in* (1.10) *and the bounds given in Theorem* 1.1.

Theorem 1.2 is proven in section 6. By eliminating the diffusion term ($D = 0$), the system of equations (1.7)–(1.9) can be reduced to a set of three, first-order ordinary differential equations. A conserved quantity enables us to reduce the system further to two equations. Thus, a phase portrait solution is easily obtained and we show that the traveling wave solution is unique. It is interesting to note that the question of existence and uniqueness for the system *with diffusion* can also be considered by examining the flow in the phase plane. In this case, the phase space has five dimensions although it is reduced to four via the conserved quantity. One of the equilibrium points is degenerate so it is difficult to determine the flow path. As such, we have elected to pursue the proof of existence by utilizing degree theory. Moreover, numerical studies [8] suggest that the traveling wave solution *with diffusion* is not unique.

The rest of the paper is organized as follows. In section 2, we show the existence of solutions to a regularized system on any finite interval via degree theory, based on maximum principles and preliminary a priori bounds. In section 3, we carry out refined a priori estimates of solutions as the length of the interval tends to infinity. In

particular, we obtain uniform estimates of the wave speed and bound it strictly inside the interval $(\frac{v}{R_f}, v)$. In section 4, utilizing the estimates of section 3, we pass to the infinite line limit of solutions and justify the validity of the boundary conditions of the limiting solutions. In section 5, we obtain further ε-independent bounds of solutions, pass to the $\varepsilon \to 0$ limit of solutions, and finish the proof of Theorem 1.1. In section 6, we prove Theorem 1.2.

2. A regularized system on finite intervals. In this section, we construct solutions to a regularized system with Dirichlet boundary conditions. In the following sections, these solutions are shown to converge to the desired traveling wave solutions of (1.7)–(1.9) as we pass to the infinite line limit and remove the regularization. We add an elliptic regularization term $\epsilon M_{\xi\xi}$ to the left side of (1.9) with $\epsilon \in (0,1)$ so that the existence problem is turned into a fixed point problem for which classical Leray–Schauder degree theory [4], [13] is available. We derive a priori bounds of solutions and compute the degree using its homotopic invariance as in [1] and [11]. The nonzero degree (equal to -1 in our case) implies the existence of a solution.

Let us first normalize the original system (1.7)–(1.9) so that K_S and K_A are scaled to one. Define $S = K_S\hat{S}$, $A = K_A\hat{A}$, $q = K_S\hat{q}$, $\gamma = \frac{K_A}{K_S}\hat{\gamma}$, and $Y = \frac{1}{K_S}\hat{Y}$. Then (1.7)–(1.9) remains the same under hat variables and parameters except that K_S and K_A are replaced by one. With no loss of generality, we also set $\hat{q} = 1$. The normalized system (without the hats) reads

$$DS_{\xi\xi} + (R_f c - v)S_\xi = MSA(1+S)^{-1}(1+A)^{-1}, \tag{2.1}$$

$$DA_{\xi\xi} + (c - v)A_\xi = \gamma MSA(1+S)^{-1}(1+A)^{-1}, \tag{2.2}$$

$$cM_\xi = -YMSA(1+S)^{-1}(1+A)^{-1} + b(M - M_0) \tag{2.3}$$

under the boundary conditions (1.10).

We propose to study the following regularized elliptic system and the associated boundary value problem on a finite interval $[-d, d]$:

$$DS_{\xi\xi} + (R_f c - v)S_\xi = R^{(1)}_{s,\varepsilon}, \tag{2.4}$$

$$DA_{\xi\xi} + (c - v)A_\xi = \gamma R^{(2)}_{s,\varepsilon}, \tag{2.5}$$

$$\varepsilon M_{\xi\xi} + cM_\xi = -YR^{(3)}_{s,\varepsilon} + b(M - M_0), \tag{2.6}$$

with $\varepsilon \in (0,1)$ and the boundary conditions

$$S(-d) = A(d) = 0,\ S(d) = S^+,\ A(-d) = A_-,\ M(\pm d) = M_0. \tag{2.7}$$

To remove the translation invariance of traveling wave solutions of an unknown speed c, we also impose the additional normalization condition

$$A(0) = \theta, \qquad \theta \in (0, A_-), \tag{2.8}$$

with θ prescribed. The modified reaction terms $R^{(i)}_s$, $i = 1, 2, 3$, are

$$\begin{aligned} R^{(1)}_{s,\varepsilon} &= \frac{|M|S|A|}{(1+\varepsilon|M|)(1+|S|)(1+|A|)}, \\ R^{(2)}_{s,\varepsilon} &= \frac{|M||S|A}{(1+\varepsilon|M|)(1+|S|)(1+|A|)}, \\ R^{(3)}_{s,\varepsilon} &= \frac{M|S||A|}{(1+\varepsilon|M|)(1+|A|)(1+|S|)}\,. \end{aligned} \tag{2.9}$$

Note that we modify the reaction terms by the factor $(1+\varepsilon|M|)$ which is crucial to obtain the first upper bound on M. The variables S, A, M, and c depend on both d and ε; however, to simplify the notation, we do not specify this dependence until it becomes necessary in the later sections.

In order to use degree theory, we consider a parametrized family of equations indexed by $\tau \in [0,1]$,

$$DS_{\xi\xi} + (R_f c - v)S_\xi = \tau R^{(1)}_{s,\varepsilon}, \tag{2.10}$$

$$DA_{\xi\xi} + (c-v)A_\xi = \gamma\tau R^{(2)}_{s,\varepsilon}, \tag{2.11}$$

$$\varepsilon M_{\xi\xi} + cM_\xi = -Y\tau R^{(3)}_{s,\varepsilon} + b(M - M_0), \tag{2.12}$$

under the boundary conditions (2.7) and the imposed condition (2.8).

Note that if $\tau = 0$, (2.10)–(2.12) under (2.7) and (2.8) has a unique solution. Without (2.8), the system is uniquely solvable for any given c, as a two-point boundary value problem for second-order ordinary differential equations. The solutions are

$$\begin{aligned} S(\xi) &= S^+ \frac{1 - e^{(-\frac{cR-v}{D})(\xi+d)}}{1 - e^{(-\frac{cR-v}{D})(2d)}}, \\ A(\xi) &= A_- \frac{1 - e^{(-\frac{c-v}{D})(\xi-d)}}{1 - e^{(\frac{c-v}{D})(2d)}}, \\ M(\xi) &= M_0. \end{aligned} \tag{2.13}$$

The extra condition $A(0) = \theta$ implies

$$\theta = A_- \frac{1 - e^{d(\frac{c-v}{D})}}{1 - e^{2d(\frac{c-v}{D})}} = A_- \frac{1}{1 + e^{d(\frac{c-v}{D})}},$$

which uniquely determines c, since the right-hand side is a monotone function of c and ranges between zero and A_-.

Next we derive a priori estimates on solutions of (2.10)–(2.12) under the boundary and normalization conditions (2.7) and (2.8) independent of $\tau \in [0,1]$.

PROPOSITION 2.1. *Let $\tau \in [0,1]$ and let (S, A, M, c) be a solution to* (2.10)–(2.12) *subject to the boundary conditions* (2.7). *Then $\forall \tau \in [0,1]$ and $\forall \xi \in [-d,d]$, we have the following inequalities:*

$$0 \le S(\xi) \le S^+, \quad 0 \le A(\xi) \le A_-, \tag{2.14}$$

$$M_0 \le M(\xi) \le \frac{1}{2\varepsilon}\left(\varepsilon M_0 + \frac{Y}{b} - 1 + \sqrt{\left(1 - \frac{Y}{b} - \varepsilon M_0\right)^2 + 4\varepsilon M_0}\right) \equiv M_{\max}, \tag{2.15}$$

$$S'(\xi) > 0, \quad A'(\xi) < 0. \tag{2.16}$$

Proof. In view of (2.10)–(2.11), both S and A satisfy the classical elliptic strong maximum principle; see, for example, [9]. The maximum and minimum of S and A are achieved at the end points. Hence, (2.14) follows from (2.7). To show that $M \ge M_0$ $\forall \xi \in [-d,d]$, suppose that $M < M_0$ at some point $\xi_0 \in (-d,d)$ for some $\tau = \tau_1 \in (0,1)$. Then, since M is continuous in τ and $\xi \in [-d,d]$, as τ varies, the minimum of M must pass through the interval $(0, M_0)$ before M becomes negative. Let us assume that at τ_1, $\min_{\xi\in[-d,d]} M \in (0, M_0)$. Hence, $\exists\ \xi_1 \in (-d,d)$ and $0 <$

$M(\xi_1) = \min_{\xi\in[-d,d]} M(\xi) < M_0$, for $\tau = \tau_1$. Evaluating (2.12) at $\xi = \xi_1$ and $\tau = \tau_1$, we have

$$-Y\tau_1 R^{(3)}_{s,\varepsilon}(\xi_1) + b(M(\xi_1) - M_0) = \varepsilon M_{\xi\xi}(\xi_1) + cM_\xi(\xi_1).$$

At the minimum, the first derivative is zero and the second derivative is nonnegative, which results in

$$-Y\tau_1 R^{(3)}_{s,\varepsilon}(\xi_1) = \varepsilon M_{\xi\xi}(\xi_1) + b(M_0 - M(\xi_1)) > 0.$$

Since $Y, \tau_1 > 0$, we have that $R^{(3)}_{s,\varepsilon}(\xi_1) < 0$. This is impossible since $M(\xi_1) > 0$. It follows that $M \geq M_0\ \forall\tau \in [0,1]$, $\xi \in [-d,d]$. By now, since $A \geq 0,\ S \geq 0,\ M \geq M_0$, we can identify $R^{(1)}_{s,\varepsilon} = R^{(2)}_{s,\varepsilon} = R^{(3)}_{s,\varepsilon} \equiv R_{s,\varepsilon}$.

To prove the upper bound in (2.15), we define $\max_{\xi\in[-d,d]} M(\xi) = M^* \geq M_0$. We only need to consider the case $M^* > M_0$. There exists $\xi^* \in (-d,d)$ such that $M(\xi^*) = M^*, M'(\xi^*) = 0, M''(\xi^*) \leq 0$. Evaluating (2.12) at $\xi = \xi^*$ implies that

$$b(M^* - M_0) \leq \left.\frac{Y\tau AS M^*}{(1+A)(1+\varepsilon M^*)(1+S)}\right|_{\xi=\xi^*} \leq \frac{YM^*}{1+\varepsilon M^*}. \tag{2.17}$$

This can be rewritten as $b(M^* - M_0)(1+\varepsilon M^*) \leq YM^*$, or

$$M^* \leq \frac{1}{2\varepsilon}\left(\varepsilon M_0 + \frac{Y}{b} - 1 + \sqrt{\left(\varepsilon M_0 + \frac{Y}{b} - 1\right)^2 + 4\varepsilon M_0}\ \right) = M_{\max}. \tag{2.18}$$

Notice that if $\frac{Y}{b} \geq 1$, then the right-hand side of (2.18) behaves like $O(\varepsilon^{-1})$ ($O(\varepsilon^{-1/2})$ if $Y = b$) as $\varepsilon \to 0$. If $\frac{Y}{b} < 1$, then in the limit $\varepsilon \to 0$, the right-hand side converges to $\frac{M_0}{1-b^{-1}Y} > M_0$. To prove (2.16), we rewrite (2.4) by multiplying both sides of the equation by $e^{\frac{(cR_f - v)\xi}{D}}$. Then, integrating from $-d$ to ξ, we obtain

$$e^{\frac{(cR_f-v)\xi}{D}} S'(\xi) = e^{\frac{-(cR_f-v)d}{D}} S'(-d) + \int_{-d}^{\xi} \frac{1}{D} e^{\frac{(cR_f-v)\xi'}{D}} R_s(\xi')d\xi'.$$

Since S is not identically constant and, as a result of the Hopf lemma, it is clear that $S'(-d) > 0$. Thus, the entire right-hand side is positive and $S' > 0$. Similarly, it can be shown that $A' < 0$. The proof of the proposition is complete. ∎

PROPOSITION 2.2. *Let $\tau \in [0,1]$ and let $d > 1$ be fixed. There exists a constant $\underline{c}$ independent of $\tau \in [0,1]$ such that the wave speed $c = c(\tau)$ in (2.10)–(2.12) satisfies*

$$\underline{c} \leq c(\tau) \leq v + \frac{D}{d}\ln\left(\frac{1}{\theta_0} - 1\right), \tag{2.19}$$

where $\theta_0 = \frac{\theta}{A_-} \in (0,1)$.

Proof. To establish the upper bound for c, we find an upper solution $\overline{A}(\xi)$ to $A(\xi)$ and use it to bound c from above. The upper solution solves

$$\begin{cases} D\overline{A}'' + (c-v)\overline{A}' = 0, & \xi \in [-d,d], \\ \overline{A}(-d) = A_-, \quad \overline{A}(d) = 0, \end{cases} \tag{2.20}$$

and is given by

$$\overline{A}(\xi) = \frac{A_-(1 - e^{-(\frac{c-v}{D})(\xi - d)})}{1 - e^{(\frac{c-v}{D})2d}}. \tag{2.21}$$

By definition, $A(\xi) \le \overline{A}(\xi)$, and in particular $A(0) \le \overline{A}(0)$. Evaluating $\overline{A}$ at $\xi = 0$ and solving for c, we have

$$c \le v + \frac{D}{d} \ln\left(\frac{A_-}{\theta} - 1\right) = v + \frac{D}{d} \ln\left(\frac{1}{\theta_0} - 1\right). \tag{2.22}$$

With (2.22), we have established the upper bound for the wave speed $c(\tau)$. To show the lower bound, we proceed similarly by defining a lower solution $\underline{A}(\xi)$ for $A(\xi)$ on $[-d, d]$ which solves

$$D\underline{A}'' + (c - v)\underline{A}' = ([S^+ - S(0, d)]H(\xi) + S(0, d))F\underline{A}, \tag{2.23}$$

where

$$F = F(\varepsilon) \equiv \frac{M_{\max}}{1 + \varepsilon M_0} \ge \frac{M}{1 + \varepsilon M}; \quad H(\xi) = \begin{cases} 1, & \xi \ge 0, \\ 0, & \xi < 0, \end{cases}$$

along with boundary and regularity conditions

$$\underline{A}(-d) = A_-, \quad \underline{A}(d) = 0, \quad \underline{A} \in C^1. \tag{2.24}$$

It follows from (2.23) that

$$\underline{A}(\xi) = \begin{cases} c_1 e^{r_1 \xi} + c_2 e^{r_2 \xi}, & \xi \in [-d, 0], \\ c_3 e^{r_3 \xi} + c_4 e^{r_4 \xi}, & \xi \in [0, d], \end{cases} \tag{2.25}$$

where $r_{1,2}$ and $r_{3,4}$ are given by

$$r_{1,2} = \frac{-(c - v) \pm \sqrt{(c - v)^2 + 4DFS(0, \tau)}}{2D}, \quad r_1 > 0, r_2 < 0; \tag{2.26}$$

$$r_{3,4} = \frac{-(c - v) \pm \sqrt{(c - v)^2 + 4DFS^+}}{2D}, \quad r_3 > 0, r_4 < 0. \tag{2.27}$$

Using boundary conditions (2.24) and the fact that $\underline{A}(0^+) = \underline{A}(0^-)$ and $\underline{A}'(0^-) = \underline{A}'(0^+)$, we solve for the constants

$$\begin{aligned} c_1 &= A_- e^{r_1 d} - c_2 e^{(r_1 - r_2)d}, \\ c_2 &= \frac{A_- e^{r_1 d}[(r_4 - r_3 e^{(r_4 - r_3)d}) - r_1(1 - e^{(r_4 - r_3)d})]}{(r_2 - r_1 e^{(r_1 - r_2)d})(1 - e^{(r_4 - r_3)d}) - (r_4 - r_3 e^{(r_4 - r_3)d})(1 - e^{(r_1 - r_2)d})}, \\ c_3 &= -c_4 e^{(r_4 - r_3)d}, \\ c_4 &= \frac{A_- e^{r_1 d} + c_2(1 - e^{(r_1 - r_2)d})}{1 - e^{(r_4 - r_3)d}}. \end{aligned} \tag{2.28}$$

The solution is

$$\underline{A}(\xi) = \begin{cases} A_- e^{r_1(\xi + d)} + c_2(e^{r_2 \xi} - e^{r_1 \xi} e^{(r_1 - r_2)d}), & \xi \in [-d, 0], \\ \left(\frac{A_- e^{r_1 d} + c_2(1 - e^{(r_1 - r_2)d})}{1 - e^{(r_4 - r_3)d}}\right)\left(e^{r_4 \xi} - e^{(r_4 - r_3)d} e^{r_3 \xi}\right), & \xi \in [0, d], \end{cases} \tag{2.29}$$

where c_2 is given in (2.28).

To find the lower bound for c, we assume there is no lower bound ($c \to -\infty$) and use the following two properties to deduce a contradiction: $\underline{A}(0) \le \theta$, $\underline{A}(0) = A_- e^{r_1 d} + c_2(1 - e^{(r_1 - r_2)d})$. Combining these with equations (2.28)–(2.29), we find that

$$\begin{aligned}\theta &\ge \underline{A}(0,\tau)\\ &= \frac{A_- e^{r_2 d}(1 - e^{(r_4 - r_3)d})(r_2 - r_1)}{(r_2 e^{(r_2 - r_1)d} - r_1)(1 - e^{(r_4 - r_3)d}) - (r_4 - r_3 e^{(r_4 - r_3)d})(e^{(r_2 - r_1)d} - 1)}.\end{aligned} \tag{2.30}$$

Fix any $\tau \in [0,1]$, and let $c \to -\infty$. We have from (2.26)–(2.27) that $r_1 \to +\infty$, $r_2 \to 0^-$, $r_3 \to +\infty$, and $r_4 \to 0^-$. It follows from (2.30) that $\lim_{c \to -\infty} \underline{A}(0,\tau) = A_- \le \theta$, contradicting the fact that $\theta \in (0, A_-)$. Since $S(0,\tau)$ is bounded between zero and S_+, it is easy to see that the limit of $A(0,\tau)$ as $c \to -\infty$ is uniform in $\tau \in [0,1]$. Hence, there exists a constant $\underline{c}$ independent of $\tau \in [0,1]$ such that $c(\tau) \ge \underline{c}$. The proof is complete. □

REMARK 2.1. *By Propositions* 2.1 *and* 2.2 *and by standard elliptic estimates, the maximum norms of the derivatives of solutions* (S', A', M') *are bounded independently of the parameter* $\tau \in [0,1]$.

Next we show the existence of solutions to (2.10)–(2.12), (2.7), and (2.8) on the bounded domain $[-d, d]$ by Leray–Schauder degree theory. The idea is to transform the system of equations into a fixed point problem. The solution of the fixed point problem solves the original system of equations ($\tau = 1$) which is topologically equivalent to a simpler system ($\tau = 0$) for which we can easily determine the degree. Thus, showing that the degree is not zero amounts to proving the existence of a solution to the original system ($\tau = 1$). The proof is similar to that given in [1]; thus, the definitions and propositions are listed but the details of the proof can be found by referring to [1].

We begin by fixing d and defining $I_d = [-d, d]$ and $X = (C^1(I_d))^3 \times R$. The set X is a Banach space equipped with the norm

$$\|(S, A, M, c)\|_X \equiv \max\left(\|S\|_{C^1(I_d)}, \|A\|_{C^1(I_d)}, \|M\|_{C^1(I_d)}, |c|\right). \tag{2.31}$$

For each element $(S, A, M, c) \in X$, we consider the unique solution $(\overline{S}, \overline{A}, \overline{M})$ of the following system indexed by $\tau \in [0,1]$:

$$D\overline{S}_{\xi\xi} + (R_f c - v)\overline{S}_\xi = \tau \frac{qSAM}{(1+S)(1+A)(1+\varepsilon M)}, \tag{2.32}$$

$$D\overline{A}_{\xi\xi} + (c - v)\overline{A}_\xi = \gamma\tau \frac{qSAM}{(1+S)(1+A)(1+\varepsilon M)}, \tag{2.33}$$

$$\varepsilon\overline{M}_{\xi\xi} + c\overline{M}_\xi - b(\overline{M} - M_0) = -Y\tau \frac{qSAM}{(1+S)(1+A)(1+\varepsilon M)} \tag{2.34}$$

subject to the boundary and normalization conditions (2.7) and (2.8). We define a compact map $K_\tau : (S, A, M, c) \in X \to (\overline{S}, \overline{A}, \overline{M}, c - \overline{A}(0) + \theta) \in X$. Every solution of (2.10)–(2.12) is a fixed point of the map K_τ given by $K_\tau[(S, A, M, c)] = (S, A, M, c)$. We define another map, $F_\tau = I - K_\tau$, where I is the identity map in X $F_\tau : (S, A, M, c) \to (S - \overline{S}, A - \overline{A}, M - \overline{M}, \overline{A}(0) - \theta)$. Every solution to (2.10)–(2.12) is a solution of $F_\tau = 0$.

The set Ω, upon which the degree is well defined, is

$$\Omega = \{(S, A, M, c) \in X \,|\, \|S\|_{C^1} \le N, \|A\|_{C^1} \le N, \|M\|_{C^1} \le N, |c| \le N\}, \tag{2.35}$$

where N is larger than the τ-independent a priori bounds of C^1 norms of (S, A, M) in Proposition 2.1 and Remark 2.1; N is also larger than the absolute value of the upper and lower bounds of $c(\tau)$ given in Proposition 2.2. The degree of F_τ at 0, or $deg(F_\tau, \overline{\Omega}, 0)$, is known to be well defined if $F_\tau(\partial\Omega) \neq 0$.

PROPOSITION 2.3. *For all* $\tau \in [0,1]$, $F_\tau(\partial\Omega) = (I - K_\tau)(\partial\Omega) \neq 0$.

Proof. See Proposition 7.3 in [1]. □

PROPOSITION 2.4. $F_\tau = (I - K_\tau)$ *satisfies,* $\forall \tau \in [0,1]$,

$$deg(F_\tau, \overline{\Omega}, 0) = deg(F_0, \overline{\Omega}, 0) = -1. \tag{2.36}$$

Proof. Equations (2.13) are the solutions denoted in this proof by (S^0, A^0, M^0). See Proposition 7.5 in [1] for more details. □

3. Further estimates on the regularized system. In this section, we obtain further estimates on solutions of the system (2.4)–(2.8) which enable us to extend the solutions to the entire real line in section 4. In particular, by carefully choosing the value of θ, we find the correct wave speed and prevent solutions from converging to the trivial solution as $d \to \infty$.

To begin, we rewrite the system of equations as

$$DS_{\xi\xi} + (R_f c - v)S_\xi = R_{s,\varepsilon}, \tag{3.1}$$

$$DA_{\xi\xi} + (c - v)A_\xi = \gamma R_{s,\varepsilon}, \tag{3.2}$$

$$\varepsilon M_{\xi\xi} + cM_\xi = -Y R_{s,\varepsilon} + b(M - M_0), \tag{3.3}$$

where $\varepsilon \in (0,1)$. The boundary conditions are

$$S(-d) = A(d) = 0,\ S(d) = S^+,\ A(-d) = A_-,\ M(\pm d) = M_0, \tag{3.4}$$

and the normalization condition is

$$A(0) = \theta, \qquad \theta \in (0, A_-). \tag{3.5}$$

The modified reaction term, $R_{s,\varepsilon}$, is

$$R_{s,\varepsilon} = \frac{MSA}{(1 + \varepsilon M)(1 + S)(1 + A)}.$$

PROPOSITION 3.1. *There exist* $\theta_0^* \in (0,1)$ *and constants* α, $d_0 > 0$ *independent of* d, *such that if* $\theta_0 \in (0, \theta_0^*)$, $\theta = \theta_0 A_-$, *and* $d \geq d_0$, *the wave speed* c *in* (3.1)–(3.3) *satisfies*

$$\frac{D}{dR_f}\ln\alpha + \frac{v}{R_f} - \frac{D}{dR_f}[\ln d + \ln S^+] \leq c \leq v + \frac{D}{d}\ln\left(\frac{1}{\theta_0} - 1\right), \tag{3.6}$$

implying in the limit $d \to \infty$

$$\frac{v}{R_f} \leq \liminf_{d\to\infty} c \leq \limsup_{d\to\infty} c \leq v. \tag{3.7}$$

Proof. In view of Proposition 2.2, we need only to establish the lower bound of (3.6). To find the lower bound, we use the same approach as in Proposition 2.2 but consider an upper solution for S (rather than A) which satisfies

$$\begin{cases} D\overline{S}'' + (cR_f - v)\overline{S}' = 0, & \xi \in [-d, d], \\ \overline{S}(-d) = 0, \quad \overline{S}(d) = S^+. \end{cases} \tag{3.8}$$

Solving this differential equation and evaluating it at $\xi = 0$, we obtain the inequality

$$c \geq \frac{v}{R_f} - \frac{D}{R_f d} \ln\left(\frac{S^+}{S(0)} - 1\right), \tag{3.9}$$

where $S(0) = S(0,d)$, due to the implicit dependence of $S(0)$ on d. To find the lower bound for c, we must bound $S(0,d)$ from below. Defining $L = \liminf_{d\to\infty} S(0,d)d$, consider the following two cases for L. First, suppose that $0 < L \leq \infty$. Then, $S(0,d) \geq \frac{\alpha}{d}$ for $\alpha \in (0, \frac{L}{2})$, and d large enough. By (3.9),

$$c \geq \frac{v}{R_f} - \frac{D}{R_f d} \ln\left[\frac{dS^+}{\alpha} - 1\right] \geq \frac{v}{R_f} - \frac{D}{dR_f}[\ln d + \ln S^+] + \frac{D}{dR_f} \ln \alpha, \tag{3.10}$$

and so, for $0 < L \leq \infty$, the proposition is proved.

Now suppose that $L = 0$; then $\forall \delta > 0$, $\exists\ d(\delta)$ such that if $d \geq d(\delta)$, $S(0,d) \leq \frac{\alpha}{d}$. Since $\limsup_{d\to\infty} c \leq v$, we can assume that $\liminf_{d\to\infty} c \leq \frac{v}{R_f}$; otherwise the proposition holds for large enough d. Therefore, there exists a sequence $\{d_j\} \to \infty$, such that $c(d_j) \to c^*$, with $c^* \leq \frac{v}{R_f}$.

Evaluating the lower solution $\underline{A}(\xi)$ in the proof of Proposition 2.2 (with $\tau = 1$, $d = d_j$) at $\xi = 0$, we have $\underline{A}(0) = A_- e^{r_1 d_j} + c_2(1 - e^{(r_1 - r_2)d_j})$, where $r_{1,2}$ and $r_{3,4}$ are given by (2.26)–(2.27) with $S(0,\tau)$ replaced by $S(0,d_j)$, and c_2 as in (2.28). It follows that

$$\begin{aligned} &\lim_{j\to\infty} r_1 = \frac{v - c^*}{D},\ \lim_{j\to\infty} r_2 = 0, \\ &\lim_{j\to\infty} r_{3,4} = \frac{(v-c^*) \pm \sqrt{(v-c^*)^2 + 4DFS^+}}{2D}. \end{aligned} \tag{3.11}$$

Using the assumption that $L = 0$,

$$\lim_{j\to\infty} r_2 d_j = \lim_{j\to\infty} \frac{-2FS(0,d_j)d_j}{(v-c) + \sqrt{(v-c)^2 + 4DFS(0,d_j)}} = 0. \tag{3.12}$$

Therefore, we infer from (3.11)–(3.12) that

$$\theta \geq \lim_{j\to\infty} \underline{A}(0) \geq A_- \frac{2}{1 + \sqrt{1 + \frac{4DFS^+}{v^2(1-\frac{1}{R_f})^2}}}. \tag{3.13}$$

Recall that $\theta = \theta_0 A_-$. We see from (3.13) that

$$\theta_0 \geq \theta_0^* \equiv \frac{2}{1 + \sqrt{1 + \frac{4DFS^+}{v^2(1-R_f^{-1})^2}}}.$$

Since $\theta_0 \in (0,1)$, we can choose $\theta_0 < \theta_0^*$ to deduce a contradiction. So for large d and $\theta_0 < \theta_0^*$, we conclude that $L > 0$. We have proven Proposition 3.1. □

COROLLARY 3.1. *If* $\liminf_{d\to\infty} c = \frac{v}{R_f}$ *and* $\theta \in (0, \theta_0^*)$, *then*

$$\liminf_{d\to\infty} S(0,d)d > 0.$$

Proof. In view of (3.12), the proof above implies that $\liminf_{d\to\infty} S(0,d)d > 0$ which is equivalent to saying that there exists a constant $\alpha = \alpha(\varepsilon) > 0$, such that for d large, $S(0,d) \geq \alpha d^{-1}$. In other words, if $S(0,d)$ tends to zero, then it goes to zero more slowly than $\frac{1}{d}$. ∎

LEMMA 3.1. *If* $\liminf_{d\to\infty} c = \frac{v}{R_f}$ *and* $\theta \in (0, \theta_0^*)$, *then the derivative* $A'(d)$ *satisfies*

$$\limsup_{d\to\infty} |A'(d)| \leq \frac{\theta}{D} \limsup_{d\to\infty} \sqrt{(v-c)^2 + \frac{4\gamma D S(0,d) M_0}{(1+S^+)(1+A_-)}}$$
$$\leq \frac{\theta}{D} \sqrt{\left(v - \frac{v}{R_f}\right)^2 + \frac{4\gamma D M_0}{1+A_-}}.$$

Proof. First, notice that $\forall \xi \in [0,d]$, we have $\gamma R_s \geq \eta_1(d) A$, where

$$\eta_1(d) = \frac{\gamma S(0,d) M_0}{(1+S^+)(1+A_-)(1+\varepsilon M_{\max})} \geq O(d^{-1}), \tag{3.14}$$

and the lower bound is given by Corollary 3.1. Now, define $\overline{A}(\xi)$ to solve

$$\begin{cases} D\overline{A}'' + (c-v)\overline{A}' - \eta_1(d)\overline{A} = 0, & \xi \in [0,d], \\ \overline{A}(0) = \theta, \quad \overline{A}(d) = 0. \end{cases} \tag{3.15}$$

Solving (3.15),

$$\overline{A}(\xi) = \theta e^{r_2 \xi} \frac{e^{(r_1-r_2)d} - e^{(r_1-r_2)\xi}}{e^{(r_1-r_2)d} - 1}, \; r_{1,2} = \frac{v - c \pm \sqrt{(v-c)^2 + 4\eta_1(d) D}}{2D}. \tag{3.16}$$

By the assumptions, Corollary 3.1, and (3.14), we see that $-(r_2 - r_1)d \to +\infty$, so $e^{(r_2-r_1)d} \to 0$ and $e^{r_2 d} \leq 1$. Therefore, by Proposition 3.1,

$$\lim_{d\to\infty} |A'(d)| \leq \lim_{d\to\infty} |\overline{A}'(d)| \leq \frac{\theta}{D} \sqrt{\left(v - \frac{v}{R_f}\right)^2 + \frac{4DM_0\gamma}{1+A_-}}. \tag{3.17}$$

The proof is complete. ∎

Note that if $\liminf_{d\to\infty} \eta_1(d) > 0$, $\lim_{d\to\infty} e^{r_2 d} = 0$. Then, $\lim_{d\to\infty} A'(d) = 0$. However, we need more results to deduce that $\eta_1(d)$, or rather $S(0,d)$, does not converge to zero as $d \to \infty$.

PROPOSITION 3.2. *If* $\theta \in (0, \theta_0^*)$, *then* $\lim_{d\to\infty} S'(-d) = \lim_{d\to\infty} A'(-d) = 0$.

Proof. Step 1. $\lim_{d\to\infty} |S'(-d)| = 0$. To prove this, we find an upper solution to S and use a resulting inequality to bound $\lim_{d\to\infty} |S'(-d)|$ by zero on both sides. To construct an upper solution, we note that on $[-d, 0]$, $R_s \geq \eta_2 S$, where we define $\eta_2 = \frac{M_0\theta}{(1+\varepsilon M_{\max})(1+A_-)(1+S^+)}$, which is a constant, independent of d. Let $\overline{S}$ solve (3.8) with the right-hand side equal to $\eta_2 \overline{S}$ and $\overline{S}(-d) = 0$, $\overline{S}(0) = S(0,d)$ for $\xi \in [-d, 0]$. Solving for $\overline{S}$, we obtain

$$\overline{S}(\xi) = S(0,d) e^{\bar{r}_1 \xi} \left(\frac{e^{(\bar{r}_1 - \bar{r}_2)d} - e^{(\bar{r}_2 - \bar{r}_1)\xi}}{e^{(\bar{r}_1 - \bar{r}_2)d} - 1} \right), \quad \xi \in [-d, 0], \tag{3.18}$$

where

$$\bar{r}_{1,2} = \frac{(v - cR_f)}{2D} \pm \frac{\sqrt{(v-cR_f)^2 + 4D\eta_2}}{2D}, \quad \bar{r}_1 > 0, \bar{r}_2 < 0.$$

From (3.18), we find $\overline{S}'(-d) = \frac{S(0,d)(\bar{r}_1-\bar{r}_2)}{e^{\bar{r}_1 d}-e^{\bar{r}_2 d}}$. Note that if $c \not\to \frac{v}{R_f}$ as $d \to \infty$, then by Proposition 3.1, $c > \frac{v}{R_f}$ for large d, and so $\bar{r}_2 < -\sqrt{\eta_2/D}$ for large d. If $c \to \frac{v}{R_f}$, as $d \to \infty$, then $\bar{r}_2 \to -\sqrt{\eta_2/D}$. In either case, $e^{\bar{r}_2 d}$ converges to zero, and $e^{\bar{r}_1 d}$ converges to infinity. Thus, $0 \le \limsup_{d\to\infty} S'(-d) \le \limsup_{d\to\infty} \overline{S}'(-d) = 0$, and the proof of Step 1 is complete.

Step 2. $\lim_{d\to\infty} A'(-d) = 0$. We infer from (3.18) that

$$\overline{S}(\xi) \le S(0,d)e^{\bar{r}_1\xi}\frac{e^{(\bar{r}_1-\bar{r}_2)d}}{e^{(\bar{r}_1-\bar{r}_2)d}-1} \le 2S^+ e^{\bar{r}_1\xi} \quad \forall \xi \in [-d,0], \tag{3.19}$$

where $\bar{r}_1$ is bounded away from zero uniformly in $d \to \infty$. Using (3.19), we find an upper bound for R_s to be $\gamma R_s \le \eta_3(d)A \ \forall \xi \in [-d,0]$, where η_3 is given in (3.20). We define a lower solution, $\underline{A}$, as the solution of

$$\begin{cases} D\underline{A}'' + (c-v)\underline{A}' - \eta_3\underline{A} = 0, \quad \eta_3(d) = \frac{2\gamma S^+ M_{\max} e^{-\bar{r}_1\frac{d}{2}}}{(1+\varepsilon M_0)(1+\theta)}. \\ \underline{A}(-d) = A_-, \quad \underline{A}\left(-\frac{d}{2}\right) = \theta \quad \forall \xi \in \left[-d, -\frac{d}{2}\right]. \end{cases} \tag{3.20}$$

We solve (3.20) to get $\underline{A}(\xi) = c_1 e^{r_1\xi} + c_2 e^{r_2\xi}$, with $r_{1,2}$ given by (2.26) with $FS(0,\tau)$ replaced by $\eta_3(d)$. Differentiating $\underline{A}(\xi)$, we find

$$\underline{A}'(-d) = r_1 A_- + \frac{(r_1-r_2)A_- + (r_2-r_1)\theta e^{-\frac{r_1 d}{2}}}{e^{(r_2-r_1)\frac{d}{2}}-1}. \tag{3.21}$$

Suppose that $c \to v$ along a subsequence $\{d_j\} \to \infty$; then $r_1 \to 0$ and $r_2 \to 0$. If, in addition $(r_2-r_1)d_j \to 0$, then $e^{\frac{(r_2-r_1)d_j}{2}} - 1 \sim (r_2-r_1)\frac{d_j}{2} + O((r_2-r_1)^2 d_j^2)$, and $\underline{A}'(-d_j) \to 0$. If $\liminf_{j\to\infty}(r_2-r_1)d_j > 0$, then the second term in (3.21) with $d = d_j$ goes to zero because its numerator converges to zero while its denominator does not. The first term clearly converges to zero. Hence, $\underline{A}'(-d_j) \to 0$ as $j \to \infty$.

Suppose now that $c \to c^* < v$ along a subsequence $d_j \to \infty$; then $r_1 \to \frac{|c^*-v|}{D}$ and $r_2 \to 0$ and also, $|r_2 - r_1| \to \beta > 0$, for a finite number β. Passing to the $d_j \to \infty$ limit in (3.21), we arrive at $\lim_{j\to\infty} \underline{A}'(-d_j) = 0$. It follows that we always have $\lim_{d\to\infty} A'(-d) = 0$. The proof is complete. □

LEMMA 3.2. *Let $\theta \in (0,\theta_0^*)$. There exists a positive constant K_1 depending only on $D, v, R_f, \gamma, M_{\max}$ such that* $\limsup_{d\to\infty} S'(d) \le K_1 S^+ \max(\theta, \sqrt{\theta})$.

Proof. In this proof, we construct a lower solution to S on half the finite domain to find an inequality for S'. To find the upper bound, we bound all the terms as $d \to \infty$. First, note that $A(\xi) \le \theta \ \forall \xi \in [0,d]$. Next, define the subsolution $\underline{S}$ as

$$\begin{cases} D\underline{S}'' + (cR_f - v)\underline{S}' - \eta_4\underline{S} = 0, \qquad \xi \in [0,d], \\ \underline{S}(0) = S(0,d), \quad \underline{S}(d) = S^+, \end{cases} \tag{3.22}$$

where $\eta_4 = \theta K$ and $K \equiv \gamma M_{\max}$ since $\gamma R_s \le \gamma\theta M_{\max} S$. Note that η_4 is a positive constant independent of d. The solution of (3.22) is $\underline{S} = c_1 e^{r_1\xi} + c_2 e^{r_2\xi}$, where $r_{1,2} = \bar{r}_{1,2}$ from Proposition 2.2 with η_2 replaced by η_4. Solving (3.22) and differentiating, it follows that

$$\underline{S}'(d) = \frac{(r_1-r_2)S(0,d)e^{r_2 d} - r_1 S^+ + r_2 e^{(r_2-r_1)d}S^+}{e^{(r_2-r_1)d}-1}, \tag{3.23}$$

where r_2 is rewritten as

$$r_2 = \frac{-4D\eta_4}{2D\cdot(v - cR_f + \sqrt{(v-cR_f)^2 + 4D\eta_4})}, \tag{3.24}$$

and $r_2 - r_1 = -\frac{1}{D}\sqrt{(v - cR_f)^2 + 4D\eta_4} \le -\frac{1}{D}2\sqrt{D\theta K}$. Thus, $e^{(r_2 - r_1)d} \to 0$ as $d \to \infty$. To find an upper bound on $S'(d)$, we must bound r_1 and r_2. If $v - cR_f \le 0$, then $|r_2| \ge \sqrt{\frac{\eta_4}{D}} = \sqrt{\frac{\theta K}{D}}$. If $v - cR_f > 0$, by Proposition 3.1, we see that $v - cR_f \le O(d^{-1} \ln d)$. Hence, we find a lower bound for r_2 in the limit $d \to \infty$ as

$$|r_2| \ge \frac{2\eta_4}{O(d^{-1} \ln d) + \sqrt{(O(d^{-1} \ln d))^2 + 4D\eta_4}} \to \sqrt{\frac{\eta_4}{D}},$$

and so $\limsup_{d\to\infty} r_2 \le -\sqrt{D^{-1}\eta_4} = -\sqrt{D^{-1}\theta K}$. Now, we see that $\lim_{d\to\infty} e^{r_2 d} \le \lim_{d\to\infty} e^{-\sqrt{D^{-1}\theta K} d} = 0$. We consider the same cases in order to bound r_1. If $v - cR_f \le 0$, then

$$r_1 \le \frac{2\eta_4}{((v + O(\frac{1}{d}))R_f - v) + \sqrt{((v + O(\frac{1}{d}))R_f - v)^2 + 4D\eta_4}} \tag{3.25}$$

by Proposition 3.1. Therefore,

$$\limsup_{d\to\infty} r_1 \le \frac{2\eta_4}{v(R_f - 1) + \sqrt{v^2(R_f - 1)^2 + 4D\eta_4}}. \tag{3.26}$$

If $v - cR_f > 0$, then by Proposition 3.1 again $v - cR_f \le O(d^{-1} \ln d)$ for large d, and so

$$r_1 \le \frac{O(d^{-1} \ln d) + \sqrt{O(d^{-2}(\ln d)^2) + 4D\eta_4}}{2D},$$

and finally,

$$\limsup_{d\to\infty} r_1 \le \sqrt{\frac{\theta K}{D}}. \tag{3.27}$$

In any case, we have the following:

$$\limsup_{d\to\infty} r_1 \le 2K_1(D, \gamma, v, R_f, M_{\max}) \max(\theta, \sqrt{\theta}) = K_1 \max(\theta, \sqrt{\theta}). \tag{3.28}$$

Combining (3.23) and (3.28), we obtain the result

$$\limsup_{d\to\infty} S'(d) \le S^+ \limsup_{d\to\infty} r_1 = K_1 S^+ \max(\theta, \sqrt{\theta}), \tag{3.29}$$

and the proof is complete. □

LEMMA 3.3. *There exist two positive constants δ_1 and δ_2 independent of d, and a positive number θ_0^{**} depending only on δ_1 and δ_2, $\theta_0^{**} \in (0, \theta_0^*)$. If $\theta_0 \in (0, \theta_0^{**})$, then*

$$\frac{v}{R_f} + \delta_1 \le \liminf_{d\to\infty} c \le \limsup_{d\to\infty} c \le v - \delta_2. \tag{3.30}$$

Proof. Combining the equations for S and A in (3.1)–(3.2) and integrating from $-d$ to d, we obtain

$$\gamma(DS'(d) + (cR_f - v)S^+ - DS'(-d)) = DA'(d) - DA'(-d) - (c - v)A_-.$$

We solve for c to get

$$c = \frac{D(A'(d) - A'(-d)) + \gamma D(S'(-d) - S'(d)) + vA_- + v\gamma S^+}{\gamma R_f S^+ + A_-}. \tag{3.31}$$

Using Proposition 2.1, we derive the inequality

$$\frac{DA'(d) - \gamma DS'(d) + vA_- + v\gamma S^+}{\gamma R_f S^+ + A_-} \le c \le \frac{-DA'(-d) + \gamma DS'(-d) + vA_- + v\gamma S^+}{\gamma R_f S^+ + A_-}.$$

Taking $d \to \infty$ and using Proposition 3.2, $\limsup_{d\to\infty} c < v$. If $\liminf_{d\to\infty} c > \frac{v}{R_f}$, then the proof is done if we define δ_i, $i = 1, 2$, to be the difference between c and $\frac{v}{R}$ and v, respectively. On the other hand, if $\liminf_{d\to\infty} c = \frac{v}{R_f}$, then by Lemmas 3.1 and 3.2, we have

$$\begin{aligned} &\liminf_{d\to\infty} c \\ &\quad \ge \frac{-\theta\sqrt{v^2(1 - R_f^{-1})^2 + \frac{4\gamma DM_0}{1+A_-}} - \gamma DK_1 S^+ \max(\theta, \theta^{\frac{1}{2}}) + vA_- + v\gamma S^+}{\gamma R_f S^+ + A_-}. \end{aligned} \tag{3.32}$$

Now, set $\delta_2 = (\gamma S^+(R_f - 1)v)/(\gamma S^+ R_f + A_-) > 0$, and

$$\delta_1 = \frac{v(1 - \frac{1}{R_f})(A_- + S^+\gamma) - \theta\sqrt{v^2(1 - R_f^{-1})^2 + \frac{4\gamma DM_0}{1+A_-}} - \gamma DK_1 S^+ \max(\theta, \theta^{\frac{1}{2}})}{\gamma S^+ R_f + A_-}.$$

Here δ_1 is the difference between the lower bound in (3.32) and $\frac{v}{R_f}$. There exists a $\theta_0^{**} \in (0, \theta_0^*)$ such that if $\theta_0 \in (0, \theta_0^{**})$, $\delta_1 > 0$. We end the proof. □

LEMMA 3.4. *There exists a positive constant, θ_0^{***}, depending on δ_1, δ_2 and less than θ_0^{**} such that if $\theta_0 \le \theta_0^{***}$, $\liminf_{d\to\infty} S(0, d) > 0$.*

Proof. Suppose $S(0, d_j) \to 0$ along a sequence $\{d_j\} \to \infty$; then $S(\xi) \to 0$, uniformly in $\xi \in [-d_j, 0]$. Proposition 3.1 says that $c_j = c(d_j)$ is uniformly bounded in d_j; (2.10)–(2.12) then imply that $(S, A, M)(\xi, d_j)$ is compact in $(C^1_{loc}(R))^3$. Up to a subsequence, we have $(S, A, M)(\xi, d_j) \to (\tilde{S}, \tilde{A}, \tilde{M})(\xi)$ as $j \to \infty$, $\tilde{S}' \ge 0$, $\tilde{A}' \le 0$, $\tilde{A}(0) = \theta$, $\tilde{S}(\xi) \equiv 0$ if $\xi \le 0$. By uniqueness of solutions to ordinary differential equations, we deduce from (2.10) that $\tilde{S} \equiv 0$ on R^1. Thus, $\tilde{A}$ is a bounded solution to the problem

$$\begin{cases} D\tilde{A}_{\xi\xi} + (c^* - v)\tilde{A}_\xi = 0, \quad \xi \in R^1, \\ \tilde{A}(0) = \theta. \end{cases} \tag{3.33}$$

As a consequence of Lemma 3.3, along a subsequence of $d_j \to \infty$, $c(d_j) \to c^* \in (\frac{v}{R_f} + \delta_1, v - \delta_2)$ as $j \to \infty$, where $\delta_i = \delta_i(D, v, R_f, \gamma, \theta_0, M_0, M_{\max}, S^+, A_-) > 0$ for $i = 1, 2$. The only bounded solution of (3.33) is $\tilde{A} \equiv \theta$.

Let us consider the equations for (S, A) on $[-d_j, d_j]$:

$$DS'' + (c_j R_f - v)S' = \frac{A}{1+A}\frac{M}{1+\varepsilon M}\frac{S}{1+S}, \tag{3.34}$$

$$DA'' + (c_j - v)A' = \gamma\frac{A}{1+A}\frac{M}{1+\varepsilon M}\frac{S}{1+S}. \tag{3.35}$$

Multiplying (3.34) by γ and subtracting (3.35) from the resulting equation, we obtain

$$\gamma D S'' - D A'' + \gamma(c_j R_f - v)S' - (c_j - v)A' = 0, \quad \xi \in [-d_j, d_j]. \tag{3.36}$$

Integrating once in ξ, we obtain

$$\begin{aligned} \gamma D S' - D A' + \gamma(c_j R_f - v)S - (c_j - v)A &= \gamma D S'(d_j) - D A'(d_j) + \gamma(c_j R_f - v)S^+, \\ &= Q(d_j) + \Gamma + o(1) \quad \forall \xi \in [-d_j, d_j], \end{aligned} \tag{3.37}$$

where $\Gamma = \lim_{j\to\infty} \gamma(c_j R_f - v)S^+$, and $Q(d_j) = \gamma D S'(d_j) - D A'(d_j) \geq 0$. It follows that

$$\gamma D \left(S' + \frac{c_j R_f - v}{D} S \right) = D A' + (c_j - v)A + \Gamma + Q(d_j) + o(1),$$

and so

$$\gamma D \left(e^{(\frac{c_j R_f - v}{D})\xi} S \right)' = e^{(\frac{c_j R_f - v}{D})\xi} \left[D A' + (c_j - v)A + \Gamma + Q(d_j) + o(1) \right]. \tag{3.38}$$

Integrating (3.38) over $[-d_j, \xi]$ gives

$$\begin{aligned} &\gamma D e^{(\frac{c_j R_f - v}{D})\xi} S = \int_{-d_j}^{\xi} e^{(\frac{c_j R_f - v}{D})\xi'} \left[D A' + (c_j - v)A + \Gamma + Q(d_j) + o(1) \right] d\xi' \\ &\geq D \int_{-d_j}^{\xi} e^{(\frac{c_j R_f - v}{D})\xi'} e^{(\frac{v - c_j}{D})\xi'} \cdot \left(e^{(\frac{c_j - v}{D})\xi'} A \right)' d\xi' + (\Gamma + o(1)) \int_{-d_j}^{\xi} e^{(\frac{c_j R_f - v}{D})\xi'} d\xi' \\ &= D e^{(\frac{c_j R_f - v}{D})\xi} A - D e^{-(\frac{c_j R_f - v}{D})d_j} A_- - c_j(R_f - 1) \int_{-d_j}^{\xi} e^{(\frac{c_j R_f - v}{D})\xi'} A(\xi') d\xi' \\ &\quad + (\Gamma + o(1)) \frac{D}{c_j R_f - v} e^{(\frac{c_j R_f - v}{D})\xi} + o(1). \end{aligned} \tag{3.39}$$

Recall that if $S(0, d_j) \to 0$ as $d_j \to \infty$, then $A(\xi, d_j) \to \theta$ and $S(\xi, d_j) \to 0$ as $d_j \to \infty$, $\forall \xi$, uniformly for a compact set of ξ. It follows that by the dominated convergence theorem

$$\begin{aligned} \lim_{d_j \to \infty} \int_{-d_j}^{\xi} e^{(\frac{c_j R_f - v}{D})\xi'} A(\xi') d\xi' &= \int_{-\infty}^{\xi} e^{(\frac{c^* R_f - v}{D})\xi'} \theta d\xi' \\ &= \theta \frac{D}{c^* R_f - v} e^{(\frac{c^* R_f - v}{D})\xi}, \end{aligned} \tag{3.40}$$

so (3.39) in the limit $d_j \to \infty$ reads

$$0 \geq D\theta_0 A_- - c^* D \theta_0 A_- \frac{(R_f - 1)}{(c^* R_f - v)} + \frac{\Gamma D}{c^* R_f - v}.$$

Substituting in $\Gamma = \gamma(c^* R_f - v)S^+$, we get $0 \geq \gamma S^+ D + D\theta_0 A_- \left[1 - \frac{c^*(R_f - 1)}{c^* R_f - v}\right]$, and finally

$$0 \geq S^+ + \gamma^{-1} \theta_0 A_- \left(\frac{c^* - v}{c^* R_f - v} \right) \geq S^+ - \gamma^{-1} \theta_0 A_- \frac{\delta_2}{R_f \delta_1}. \tag{3.41}$$

There exists $\theta_0^{***} \in (0, \theta_0^{**})$ depending on δ_1 and δ_2 such that the right-hand side of (3.41) is strictly positive. This implies a contradiction. Therefore, $S(0,d)$ does not tend to zero as $d \to \infty$. ∎

PROPOSITION 3.3. *If $\theta \in (0, \theta_0^{***})$, then $\lim_{d\to\infty} A'(d) = \lim_{d\to\infty} S'(d) = 0$.*

Proof. Let us follow the proof of Lemma 3.1 until the derivative of $\overline{A}$ at $\xi = d$ is found to be

$$\overline{A}'(d) = \frac{\theta e^{r_2 d}(r_2 - r_1)}{1 - e^{(r_2 - r_1)d}}, \tag{3.42}$$

where $|r_2 - r_1|$ is bounded. Now by Lemma 3.4, we can improve (3.14) so that $\liminf_{d\to\infty} \eta_1(d) > 0$. Hence, $v - c$ is bounded strictly away from zero as $d \to \infty$ by Lemma 3.3, and so $\lim_{d\to\infty} e^{r_2 d} = 0$ and $\lim_{d\to\infty} A'(d) = 0$.

For the second part of the proposition, we consider $S(\xi, d)$ over $\left[\frac{d}{2}, d\right]$. First, by (3.16),

$$A(\xi) \le \theta e^{\bar{r}_2 \frac{d}{2}} \quad \forall \xi \in \left[\frac{d}{2}, d\right], \quad \bar{r}_2 = \frac{v - c - \sqrt{(v-c)^2 + 4\eta_1(d)D}}{2D},$$

and $\limsup_{d\to\infty} \bar{r}_2 < 0$. Hence, $A(\xi, d)$ goes to zero exponentially fast in d.

Let $\eta_5(d) = M_{\max}\theta e^{\bar{r}_2 \frac{d}{2}}$, so that $R_s \le \eta_5 S$. Define a subsolution $\underline{S}$ on $[\frac{d}{2}, d]$ as in (3.22) with η_2 replaced by η_5 and $\underline{S}(\frac{d}{2}) = S(0,d)$, $\underline{S}(d) = S^+$. Solving for $\underline{S}$, taking the derivative, and evaluating it at $\xi = d$, we get

$$\underline{S}'(d) = \frac{(r_1 - r_2)S(0,d)e^{\frac{r_2 d}{2}} - r_1 S^+ + r_2 S^+ e^{(r_2 - r_1)\frac{d}{2}}}{e^{(r_2 - r_1)\frac{d}{2}} - 1} \ge S'(d) \ge 0. \tag{3.43}$$

Since c is bounded away from $\frac{v}{R_f}$ uniformly in d,

$$\lim_{d\to\infty} e^{(r_2 - r_1)\frac{d}{2}} = 0, \ \lim_{d\to\infty} e^{r_2 \frac{d}{2}} = 0, \text{ and } \lim_{d\to\infty} \eta_5(d) = 0.$$

In addition, $\lim_{d\to\infty} r_1 = 0$, and we deduce immediately that $\lim_{d\to\infty} S'(d) = 0$. The proof is complete. ∎

PROPOSITION 3.4.

$$\lim_{d\to\infty} c(d) = v\frac{(A_- + \gamma s^+)}{(A_- + \gamma R_f S^+)} = c. \tag{3.44}$$

Proof. The limit follows from (3.31) in the proof of Lemma 3.3 and Propositions 3.2 and 3.3. ∎

COROLLARY 3.2. *There exist positive constants β_1 and β_2 independent of d such that*

$$A(\xi, d) \le \theta e^{-\beta_1 \xi}, \quad \beta_1 > 0, \quad \forall \xi \in [0, d], \tag{3.45}$$

$$S(\xi, d) \le 2S^+ e^{\beta_2 \xi}, \quad \beta_2 > 0, \quad \forall \xi \in [-d, 0]. \tag{3.46}$$

Proof. To prove (3.45), we consider (3.16) from Lemma 3.1, defined on $[0, d]$. It is clear that $A(\xi, d) \le \theta e^{r_2 \xi}$, $\forall \xi \in [0, d]$. Note that $\limsup_{d\to\infty} r_2 < 0$. Letting $0 < \beta_1 \le -\limsup_{d\to\infty} r_2$ with β_1 bounded away from zero uniformly in d, we have shown (3.45). To prove (3.46), we use (3.19): $S(\xi, d) \le 2S^+ e^{\bar{r}_1 \xi}$, $\forall \xi \in [-d, 0]$. We see that $\liminf_{d\to\infty} \bar{r}_1 > 0$. Choosing $0 < \beta_2 \le \liminf_{d\to\infty} \bar{r}_1$, we have shown (3.46). ∎

4. Solutions of a regularized system on the real line. For a fixed $\varepsilon > 0$, we are interested in taking the limit $d \to \infty$. The ε dependence of solutions will not be specified until the next section. By Propositions 3.1–3.3, we know that the d dependent solutions denoted by (S_d, A_d, M_d) are compact in $(C^1_{loc}(R))^3$. So up to a subsequence in d, $(S_d, A_d, M_d) \to (S, A, M)$ uniformly on any compact set of ξ and the limiting system is

$$DS_{\xi\xi} + (R_f c - v)S_\xi = R_s, \tag{4.1}$$

$$DA_{\xi\xi} + (c - v)A_\xi = \gamma R_s, \tag{4.2}$$

$$\varepsilon M_{\xi\xi} + cM_\xi = -YR_s + b(M - M_0), \tag{4.3}$$

with boundary conditions

$$A(0) = \theta_0 A_-, \;\; \theta_0 \in (0, \theta_0^{***}), \;\; S(0) = S(0, \varepsilon) > 0. \tag{4.4}$$

Moreover, we have the following bounds:

$$0 \le S(\xi) \le S^+, \quad 0 \le A(\xi) \le A_-, \quad M_0 \le M(\xi) \le M_{\max}, \tag{4.5}$$

$$A'(\xi) \le 0, \quad S'(\xi) \ge 0. \tag{4.6}$$

Note that (4.4) is due to Lemma 3.4.

Corollary 3.2 holds for the limiting functions A and S. In particular,

$$\lim_{\xi \to +\infty} A(\xi) = 0, \quad \lim_{\xi \to -\infty} S(\xi) = 0.$$

Monotonicity of A and S in (4.6) implies that

$$\lim_{\xi \to -\infty} A(\xi) = \tilde{A}_- \quad \text{for } \tilde{A}_- \in (0, A_-], \tag{4.7}$$

$$\lim_{\xi \to +\infty} S(\xi) = \tilde{S}^+ \quad \text{for } \tilde{S}^+ \in (0, S^+]. \tag{4.8}$$

LEMMA 4.1. *For $\varepsilon \in (0,1)$, $\lim_{\xi \to \pm\infty} M(\xi) = M_0$.*

Proof. Let $m = M - M_0$; then $m \ge 0$ and $m \not\equiv 0$ (if $m \equiv 0$ by (4.3) $AS \equiv 0$ which contradicts (4.4) at $\xi = 0$). Therefore, there exists a ξ_1 such that $(M - M_0)(\xi_1) = m(\xi_1) > 0$. On $[\xi_1, \infty)$ by inequalities (4.5) and Corollary 3.2,

$$\frac{YSAM}{(1+S)(1+A)(1+\varepsilon M)} \le y_0 Y M_{\max} \theta e^{-\beta_1 \xi},$$

where $y_0 > 1$ is to be chosen and

$$\beta_1 \le \frac{c - v + \sqrt{(v-c)^2 + 4\eta_1(d)D}}{2D}, \;\; \eta_1(d) > 0.$$

Define $\bar{m}$ as a solution of $\varepsilon \overline{m}_{\xi\xi} + c\overline{m}_\xi - b\overline{m} = -y_0 Y M_{\max} \theta e^{-\beta_1 \xi}$ with $\overline{m}(\xi_1) \ge m(\xi_1) > 0$, for $\xi \in [\xi_1, \infty)$. Solving for a positive solution $\overline{m}$, we obtain

$$\overline{m} = -y_0 \frac{Y M_{\max} \theta}{(\varepsilon \beta_1^2 - c\beta_1 - b)} e^{-\beta_1 \xi}, \tag{4.9}$$

where $\beta_1 > 0$, $\varepsilon\beta_1^2 - c\beta_1 - b < 0$ for $\varepsilon \in (0,1)$ and $\beta_1 < \sqrt{b}$. So indeed $\bar{m} > 0$. We choose y_0 such that $\overline{m}(\xi_1) \geq m(\xi_1)$. By the maximum principle, $m(\xi) \leq \overline{m}(\xi)$. Since $\lim_{\xi\to+\infty} \overline{m}(\xi) = 0$, we find that $\lim_{\xi\to+\infty} m(\xi) = 0$ and $\lim_{\xi\to+\infty} M(\xi) = M_0$. A similar argument shows that $\lim_{\xi\to+\infty} M(\xi) = M_0$. □

LEMMA 4.2. *Consider M_d for $\xi \in [-d,d]$. Then $\lim_{d\to\infty} M_d'(\pm d) = 0$.*

Proof. This proof is similar to the proof of Lemma 4.1. We define $m_d = M_d - M_0$ on $[-d,d]$ and then construct an upper solution to m_d which solves

$$\begin{cases} \varepsilon\overline{m}_{d,\xi\xi} + c\overline{m}_{d,\xi} - b\overline{m}_d = -YM_{\max}\theta e^{-\beta_1\xi}, \\ \overline{m}_d(0) = m_d(0), \quad \overline{m}_d(d) = 0, \quad \xi \in [0,d], \end{cases} \tag{4.10}$$

where $c = c(d)$. We solve for $\overline{m}_d$ and find the derivative at $\xi = d$. Taking the limit as $d \to \infty$, we show that $\lim_{d\to+\infty} \overline{m'}_d(d) = 0$. It follows that $\lim_{d\to+\infty} |m_d'(d)| = \lim_{d\to+\infty} |(M_d - M_0)'(d)| = \lim_{d\to+\infty} |M_d'(d)| = 0$. A similar argument shows that $\lim_{d\to+\infty} M_d'(-d) = 0$. □

LEMMA 4.3. *There exists four constants k_1, k_2, k_3, and k_4 depending on Y, $M_{\max}$, θ, ε, b, c, R, v, and D such that*

$$(M - M_0)(\xi) \leq k_1 e^{-\sigma_1\xi} + k_2 e^{-\sigma_2\xi}, \quad \sigma_1 > 0,\ \sigma_2 > 0,\ \xi \geq 0, \tag{4.11}$$

$$(M - M_0)(\xi) \leq k_3 e^{\sigma_3\xi} + k_4 e^{\sigma_4\xi}, \quad \sigma_3 > 0,\ \sigma_4 > 0,\ \xi \leq 0, \tag{4.12}$$

where σ_i is a constant for $i = 1,2,3,4$.

Proof. The proof of (4.11) relies on bounding the upper solution $\overline{m}_d$ found in Lemma 4.2. The proof of (4.12) is similar. □

THEOREM 4.1. *For $\varepsilon \in (0,1)$, $\exists$ a smooth solution (S,A,M,c) solving the system (4.1)–(4.3) and the boundary conditions (4.4)–(4.6). Moreover, $M_0 < M \leq M_{\max}$ and $\lim_{\xi\to\pm\infty} M(\xi) = M_0$.*

Proof. We have shown that there exist solutions on the real line and that M reaches its limits at the spatial infinities. Now, we only need to show that the boundary conditions on S, A hold in the limit $d \to \infty$; i.e., we want to show that $A_- = \tilde{A}_-$ and $S^+ = \tilde{S}^+$. To that end, we multiply γ to (2.4), subtract (2.5), and integrate over $[-d,d]$. Taking the limit $d \to \infty$, using Propositions 3.2 and 3.3 and (4.7)–(4.8), we solve for c:

$$c = v\frac{(\tilde{A}_- + \gamma\tilde{S}^+)}{(\tilde{A}_- + \gamma\tilde{S}^+R_f)}. \tag{4.13}$$

Proposition 3.4 says that $c = v\frac{(A_- + \gamma S^+)}{(A_- + \gamma S^+ R_f)}$. Therefore, it is clear that $\frac{\tilde{S}^+}{S^+} = \frac{\tilde{A}_-}{A_-}$.

Multiplying (2.4) by Y, adding it to (2.6), and integrating over $[-d,d]$,

$$YD[S_d'(d) - S_d'(-d)] + Y(c_dR - v)S^+ + \varepsilon[M_d'(d) - M_d'(-d)] = b\int_{-d}^{d} (M_d - M_0)(\xi)d\xi.$$

By Propositions 3.2 and 3.3 and Lemmas 4.1 and 4.2, in the limit $d \to \infty$,

$$Y(cR_f - v)S^+ = b\lim_{d\to\infty}\int_{-d}^{d}(M_d - M_0)(\xi)d\xi = b\int_{-\infty}^{\infty}(M - M_0)(\xi)d\xi. \tag{4.14}$$

Multiplying (4.1) by Y, adding it to (4.3), integrating the resulting equation from $\xi = -\infty$ to $\xi = +\infty$, and using boundary conditions as well as the decay of the limiting functions at infinities, we end up with

$$Y(cR_f - v)\tilde{S}^+ = b\int_{-\infty}^{\infty}(M - M_0)(\xi)d\xi. \tag{4.15}$$

Combining (4.14)–(4.15), we find that $S^+ = \tilde{S}^+$, and similarly $A_- = \tilde{A}_-$. Thus, we have a solution (S, A, M) for the system (4.1)–(4.3) satisfying $S(\infty) = S^+$, $S(-\infty) = 0$, $A(\infty) = 0$, and $A(-\infty) = A_-$. The fact that $M(\xi) > M_0$ follows from the maximum principle. The proof is complete. □

5. Traveling waves on the real line. Through results of the last section, we have established the existence of the regularized solutions denoted by $(S_\varepsilon, A_\varepsilon, M_\varepsilon, c_\varepsilon)$ over the real line. We show that they converge to the desired traveling wave solutions as $\varepsilon \to 0$. We have found $c_\varepsilon = c$ independent of ε. However, we improve the previous bounds on $S_\varepsilon, A_\varepsilon, M_\varepsilon$ and their derivatives so that they are independent of ε. Moreover, we show that the boundary conditions are valid for limiting functions (S, A, M) as $\xi \to \infty$. The first step is to establish an upper bound on M_ε independent of ε.

LEMMA 5.1. *For M_ε of the system* (4.1)–(4.3)*, M_ε satisfies*

$$M_0 < M_\varepsilon \le M_0 + Y\left(R_f - \frac{v}{c}\right) S^+ \quad \forall \xi \in R^1. \tag{5.1}$$

Proof. As mentioned, $\lim_{\varepsilon \to 0} M_\varepsilon = M \not\equiv M_0$. By Proposition 2.1 and Theorem 4.1, we have that $M_0 < M_\varepsilon$. Therefore, $\exists\ \xi_1 \in (-\infty, \infty)$ such that $M_\varepsilon(\xi_1) = \sup_{\xi \in R^1} M_\varepsilon(\xi) = \overline{M}_\varepsilon > M_0$. Multiplying Y by (4.1), adding it to (4.3), and integrating from ξ_1 to ∞, we obtain

$$-Y D S'_\varepsilon(\xi_1) + Y(cR_f - v)(S^+ - S_\varepsilon(\xi_1)) + c(M_0 - \overline{M}_\varepsilon) = b \int_{\xi_1}^{\infty} (M_\varepsilon - M_0)(\xi) d\xi.$$

Note that $M'_\varepsilon(\xi_1) = 0$ and $\int_{\xi_1}^{+\infty} (M_\varepsilon - M_0) d\xi \ge 0$. Solving for $\overline{M}_\varepsilon$,

$$\overline{M}_\varepsilon \le M_0 + Y\left(R_f - \frac{v}{c}\right) S^+. \tag{5.2}$$

It is clear that $M_0 < M_\varepsilon(\xi) \le \overline{M}_\varepsilon \le M_0 + Y(R_f - v/c)S^+$. Substituting in the expression for c in Proposition 3.4, we obtain the bound given in the main theorem. The proof of the lemma is complete. □

Proof of the main theorem. First by Lemma 5.1, $(S_\varepsilon, A_\varepsilon, M_\varepsilon) \to (S, A, M)$ in C^1_{loc}. We want to show that S decays to zero as $\xi \to -\infty$ (as we showed previously in Corollary 3.2). We see that the upper bound for S_ε on $\xi \le 0$ depends on r_2 which in turn depends on $\eta_2(d, \varepsilon)$. We recall from Proposition 3.2 that η_2 is a lower bound for R_s/S_ε. Therefore, we can improve this lower bound by substituting in the upper bound for M_ε as given in Lemma 5.1. Then we have for $\varepsilon \in [0, 1]$,

$$\eta = \frac{M_0 \theta}{(1 + M_0 + Y(R_f - \frac{v}{c})S^+)(1 + S^+)(1 + A_-)} \le \eta_2^\varepsilon \le \frac{R_s^\varepsilon}{S_\varepsilon}. \tag{5.3}$$

We establish an upper bound for S_ε on the negative real line by following the proof of Corollary 3.2. In this way, we find a bound which is independent of ε and as such, $S_\varepsilon \to S$ and in the limit $\xi \to -\infty$ decays to zero.

We must also establish that $\liminf_{\varepsilon \to 0} S_\varepsilon(0) > 0$ so that the derivatives for S, A tend to zero at infinity. Following the proof of Lemma 3.3, we can derive the analogous inequality

$$0 \ge S^+ - \frac{\theta_0 \delta_2}{\gamma \delta_1}. \tag{5.4}$$

Since all these constants are independent of ε, θ_0 can be chosen small enough so that the right-hand side is positive without being dependent on ε. Thus the bound holds, and so $S(0) > 0$.

As such, it is easy to deduce that the second part of Corollary 3.2 can be reproduced independent of ε. The first inequality of Corollary 3.2 follows with β_1 independent of ε. Given the results of Lemma 5.1, and because $S(0) > 0$, we have that $\eta_1 > 0$ independent of ε. Therefore, $\lim_{\xi\to-\infty} S = 0$. Similarly, as $\xi \to \infty$, A decays to zero.

Next, we reproduce Lemma 4.1 to find the decay properties of M near infinity. We begin by improving the upper bound of R_s^ε in Lemma 4.1. We see that by Lemma 5.1,

$$R_s^\varepsilon \le \left(M_0 + Y\left(R_f - \frac{v}{c}\right)S^+\right)\theta e^{-\beta_1\xi}. \tag{5.5}$$

We construct an upper solution $\overline{m}_\varepsilon$ to $m_\varepsilon = M_\varepsilon - M_0$ on $[\xi_1, \infty]$ which solves

$$\begin{aligned} &\varepsilon\overline{m}_\varepsilon'' + c\overline{m}_\varepsilon' - b\overline{m}_\varepsilon = -Y\left(M_0 + Y\left(R_f - \frac{v}{c}\right)S^+\right)\theta e^{-\beta_1\xi}, \\ &\overline{m}_\varepsilon(\xi_1) \ge m_\varepsilon(\xi_1) > 0. \end{aligned} \tag{5.6}$$

The solution to (5.6) is

$$\overline{m}_\varepsilon = -y_1\frac{Y(M_0 + Y(R_f - \frac{v}{c})S^+)\theta e^{-\beta_1\xi}}{\varepsilon\beta_1^2 - c\beta_1 - b}, \tag{5.7}$$

where $0 < \beta_1 \le \frac{c-v+\sqrt{(v-c)^2+4D\eta_1}}{2D}$ and y_1 is a positive constant chosen so that the inequality in (5.6) holds at ξ_1. In the limit as $\varepsilon \to 0$,

$$\overline{m}_\varepsilon \to y_1\frac{Y(M_0 + Y(R_f - \frac{v}{c})S^+)\theta e^{-\beta_1\xi}}{(c\beta_1 + b)} \equiv \overline{m}_0. \tag{5.8}$$

Thus, in the limit $\xi \to +\infty$, we have that $\overline{m}_0 \to 0$, which implies that $m(\xi) \to 0$ as $\xi \to +\infty$ or $\lim_{\xi\to+\infty} M(\xi) = M_0$. Similarly, $\lim_{\xi\to-\infty} M(\xi) = M_0$.

Justifying the limits $\lim_{\xi\to-\infty} A = A_-$ and $\lim_{\xi\to+\infty} S = S^+$ as in the proof of Theorem 4.1, we have shown that the limiting functions (S, A, M, c) are traveling wave solutions satisfying all the boundary conditions.

Finally, we establish the strict inequalities for the wave profiles. By the strong elliptic maximum principle, $0 < S(\xi) < S^+$, and $0 < A(\xi) < A_-$ for any finite ξ. If $M(\xi_1) = M_0$, for a finite ξ_1, then $M'(\xi_1) = 0$. The M equation evaluated at $\xi = \xi_1$ cannot hold thanks to $S(\xi_1)A(\xi_1) > 0$. Hence, $M(\xi) > M_0$ for any finite ξ.

If $A'(\xi_2) = 0$ for a finite ξ_2, then by the A equation evaluated at $\xi = \xi_2$, we see that $A''(\xi_2) = \gamma R_s(\xi_2) > 0$. It follows that ξ_2 is a local minimal point, which contradicts the fact that A is monotone decreasing. Hence, $A' < 0$ for any finite ξ. Similarly, $S' > 0$, for any finite ξ. The proof of the main theorem is complete.

6. Existence of traveling waves in the zero-diffusion model. In this section, we prove Theorem 1.2. We have already shown that traveling wave solutions exist to (1.1)–(1.3) for $D > 0$. For the case when $D = 0$, the equations are reduced to first order. Using the conserved quantity, $\gamma R_f S - A$, and defining the new variables, $u = \gamma R_f S - A$ and $w = \gamma S - A$, (1.1)–(1.3) are transformed into conservative form

$$u_t + vw_x = 0, \tag{6.1}$$

$$w_t + v/R_f((R+1)w-u)_x = \varepsilon(u-R_f w)(u-w)/G, \tag{6.2}$$

$$M_t - b(M-M_0) = \frac{YM(u-R_f w)(u-w)}{\gamma(R-1)^2 K_S K_A G}, \tag{6.3}$$

where $\varepsilon = R_f^{-1}(R_f-1)^{-1}(K_A K_S)^{-1}$, and $G(A,S) = (1+K_A^{-1}A)(1+K_S^{-1}S)$. When rewritten in the traveling wave variable, ξ, (6.1) becomes $u_\xi = (v/c)w_\xi$. This relationship restricts the flow in the phase space to two-dimensional planes. The two-by-two dynamical system is then

$$w_\xi = -\frac{R_f\varepsilon(A_- + \gamma R_f S^+)(w-\gamma S)(w+A_-)M}{v(A_- + \gamma S^+)G},$$
$$M_\xi = \frac{b(A_- + \gamma R_f S^+)(M-M_0)}{v(A_- + \gamma S^+)} + \frac{YA_- S^+(A_- + \gamma R_f S^+)(w-\gamma S)(w+A_-)M}{v(A_- + \gamma S^+)^3 K_S K_A G}.$$

There are two equilibrium points in the phase plane (w, M): $(\gamma S^+, M_0)$ and $(-A_-, M_0)$. The eigenvalues governing the flow near $(\gamma S^+, M_0)$ are

$$\lambda_1 = \frac{b(A_- + \gamma R_f S^+)}{v(A_- + \gamma S^+)} > 0 \quad \text{and} \quad \lambda_2 = \frac{-R_f\varepsilon(A_- + \gamma R_f S^+)}{v} < 0.$$

The eigenvalues for the flow near $(-A_-, M_0)$ are both positive. Thus, $(\gamma S^+, M_0)$ is a saddle point and $(-A_-, M_0)$ is an unstable node. In order to have a traveling wave solution, at least one of the unstable manifolds emanating from the point $(-A_-, M_0)$ must intersect the two-dimensional stable manifold of $(\gamma S^+, M_0)$. It is straightforward to show that there is a unique path between the two equilibrium points. Thus, the proof of Theorem 1.2 is complete.

Acknowledgments. The authors would like to thank A. Valocchi and S. Oya for communicating their work in progress and the preprint of their paper [8]. J. X. Xin would like to thank Y. Oono for kindly introducing him to the biodegradation model studied in this paper. Both authors wish to thank J. Hyman and B. Travis of the Los Alamos National Laboratory for their interest in this work. The authors also wish to thank M. Brusseau and L. Xie for their constructive comments.

REFERENCES

[1] H. Berestycki, B. Nicolaenko, and B. Sheurer, *Traveling wave solutions to combustion models and their singular limits*, SIAM J. Math. Anal., 16 (1985), pp. 1207–1242.

[2] R. Borden and P. Bedient, *Transport of dissolved hydrocarbons influenced by oxygen-limited biodegradation*, Water Resources Res., 22 (1986), pp. 1973–1982.

[3] C. Chiang, C. Dawson, and M. Wheeler, *Modeling of in-situ bioremediation of organic compounds in groundwater*, Transport in Porous Media, 6 (1991), pp. 667–702.

[4] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd edition, Springer-Verlag, New York, 1983.

[5] F. J. Molz, M. A. Widdowson, and L. D. Benefield, *Simulation of microbial growth dynamics coupled to nutrient and oxygen transport in porous media*, Water Resources Res., 22 (1986), pp. 1207–1216.

[6] National Research Council, *In Situ Bioremediation, When Does It Work?*, Water Science and Technology Board, National Academy Press, Washington, D.C., 1993.

[7] J. Odencrantz, A. Valocchi, and B. Rittman, *Modeling the interaction of sorption and biodegradation on transport in ground water in situ bioremediation systems*, in Proceedings of the 1993 Ground Water Modeling Conference, E. Poeter, S. Ashlock, and J. Proud, eds., International Ground Water Modeling Center, Golden, CO, 1993, pp. 2-3–2-12.

[8] S. Oya and A. Valocchi, *Characterization of traveling waves and analytical estimation of pollutant removal in one-dimensional subsurface bioremediation modeling*, Water Resources Res., 33 (1997), pp. 1117–1127.
[9] M. Protter and H. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[10] A. Valocchi, *Traveling Wave Behavior During Subsurface Transport of Biologically Reactive Contaminants: Implications for In Situ Bioremediation*, private communication, Dept. of Civil Engrg., Univ. of Illinois, Urbana, IL, 1995.
[11] X. Xin, *Existence and uniqueness of traveling waves in a reaction-diffusion equation with combustion nonlinearity*, Indiana Univ. Math J., 40 (1991), pp. 985–1008.
[12] J. X. Xin, *Existence of multidimensional traveling waves in transport of reactive solutes through periodic porous media*, Arch. Rat. Mech. Anal., 128 (1994), pp. 75–103.
[13] E. Zeidler, *Nonlinear Functional Analysis*, Springer-Verlag, New York, 1984.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 95–117

SHOCK WAVES FOR A MODEL SYSTEM OF THE RADIATING GAS*

SHUICHI KAWASHIMA† AND SHINYA NISHIBATA‡

Abstract. This paper is concerned with the existence and the asymptotic stability of traveling waves for a model system derived from approximating the one-dimensional system of the radiating gas. We show the existence of smooth or discontinuous traveling waves and also prove the uniqueness of these traveling waves under the entropy condition, in the class of piecewise smooth functions with the first kind discontinuities. Furthermore, we show that the C^3-smooth traveling waves are asymptotically stable and that the rate of convergence toward these waves is $t^{-1/4}$, which looks optimal. The proof of stability is given by applying the standard energy method to the integrated equation of the original one.

Key words. shock wave, traveling wave, radiating gas, asymptotic stability, energy method

AMS subject classifications. 35B35, 35Q35, 76N15

PII. S0036141097322169

1. Introduction.

Consider the system of equations

$$u_t + uu_x + q_x = 0 \tag{1.1a}$$

and

$$-q_{xx} + q + u_x = 0 \tag{1.1b}$$

for $x \in \mathbf{R}$ and $t \geq 0$. The first equation is a hyperbolic conservation law and the second is an elliptic equation. As the second equation is a linear elliptic equation, we express q in terms of u formally as

$$q = -Ku_x, \tag{1.2}$$

where K is the inverse of the operator $-\frac{d^2}{dx^2} + 1$ and has the expression

$$(Kf)(x) = \frac{1}{2}\int_{-\infty}^{\infty} e^{-|x-y|} f(y)dy. \tag{1.3}$$

Since (1.2), we have formally that

$$q_x = -Ku_{xx} = u - Ku. \tag{1.4}$$

Thus we see that (1.1) is formally equivalent to the following:

$$u_t + uu_x + u - Ku = 0 \tag{1.5a}$$

*Received by the editors June 2, 1997; accepted for publication December 11, 1997; published electronically October 14, 1998.
http://www.siam.org/journals/sima/30-1/32216.html
†Graduate School of Mathematics, Kyushu University, Fukuoka 812-8581, Japan (kawashim@math.kyushu-u.ac.jp).
‡Department of Mathematics, Fukuoka Institute of Technology, Fukuoka 811-0295, Japan (shinya@fit.ac.jp).

and

$$q = -Ku_x. \tag{1.5b}$$

We are concerned with the traveling wave solution of (1.1), which is expressed in the form $(u, q)(x, t) = (U, Q)(\xi)(\xi = x - st)$, where s is a constant called the speed of the traveling wave. We assume that the traveling wave connects the asymptotic state u_+ and u_-, i.e.,

$$u_\pm = \lim_{\xi \to \pm\infty} U(\xi). \tag{1.6}$$

Substituting $(u, q)(x, t) = (U, Q)(\xi)$ in (1.1), we have

$$-sU^{'} + UU^{'} + Q^{'} = 0 \tag{1.7a}$$

and

$$-Q^{''} + Q + U^{'} = 0. \tag{1.7b}$$

This system of autonomous ordinary differential equations (ODEs) does not have classical solutions satisfying (1.6) when the shock $|u_+ - u_-|$ is large (see section 2). So we must look for the solutions in weak sense, which is defined as $(U, Q)(\xi)$ satisfying the following integral equations.

DEFINITION 1.1. *We define an admissible traveling wave $(U, Q)(\xi)$ as a function $(U, Q) \in L^\infty$ which satisfies*

(D.1a)
$$\int_{-\infty}^{+\infty} -s|U-k|\varphi_\xi + sign(U-k)\left(\frac{1}{2}U^2 - \frac{1}{2}k^2\right)\varphi_\xi - sign(U-k)(U-KU)\varphi d\xi \geq 0$$

and

$$\int_{-\infty}^{+\infty} Q\psi d\xi = \int_{-\infty}^{+\infty} (KU)\psi_\xi d\xi \tag{D.1b}$$

for arbitrary $\varphi \in C_0^\infty$ with $\varphi(\xi) \geq 0$, arbitrary $\psi \in \mathcal{S}$, and arbitrary real number k. Here the function sign is defined as

$$sign(x) = \begin{cases} -1 & for \;\; x < 0, \\ 0 & for \;\; x = 0, \\ 1 & for \;\; x > 0. \end{cases} \tag{1.8}$$

Here we note that this definition corresponds to the equivalent system (1.5). In order to make the above definition of the admissible traveling waves clear, we introduce different definitions of traveling waves here.

DEFINITION 1.2. *We define a traveling wave $(U, Q)(\xi)$ as a function $(U, Q) \in L^\infty$ which satisfies*

$$\int_{-\infty}^{+\infty} \left(-sU + \frac{1}{2}U^2\right)\varphi_\xi - (U - KU)\varphi d\xi = 0 \tag{D.2}$$

and the equation (D.1b) *for arbitrary $\varphi \in C_0^\infty$ and arbitrary $\psi \in \mathcal{S}$.*

DEFINITION 1.3. *We define a traveling wave* $(U,Q)(\xi)$ *as a function* $(U,Q)\in L^\infty$ *which satisfies*

$$\int_{-\infty}^{+\infty}\left(-sU+\frac{1}{2}U^2+Q\right)\varphi_\xi d\xi=0 \tag{D.3a}$$

and

$$\int_{-\infty}^{+\infty}Q(-\psi_{\xi\xi}+\psi)-U\psi_\xi d\xi=0 \tag{D.3b}$$

for arbitrary $\varphi\in C_0^\infty$ *and arbitrary* $\psi\in\mathcal{S}$.

Obviously, Definitions 1.2 and 1.3 correspond to the systems (1.5) and (1.1), respectively.

When the traveling wave $(U,Q)(\xi)$ has discontinuities of the first kind, we denote the right state of each discontinuity by (u_r,q_r) and the left state by (u_l,q_l), respectively. Also, we denote by (u_r',q_r') and (u_l',q_l') the right and the left states of $(U',Q')(\xi)$, respectively. The following proposition makes the relationship among definitions of traveling waves clear.

PROPOSITION 1.4. (i) *Definitions of traveling waves in Definitions* 1.2 *and* 1.3 *are equivalent to each other. If the traveling wave* $(U,Q)(\xi)$, $\xi=x-st$, *is a piecewise smooth function with the first kind discontinuities, it satisfies the Rankine–Hugoniot condition*

$$s(u_r-u_l)=\frac{1}{2}({u_r}^2-{u_l}^2)\left(i.e.,\quad s=\frac{u_r+u_l}{2}\right), \tag{1.9a}$$

$$q_r=q_l, \tag{1.9b}$$

and

$$u_r-q_r^{'}=u_l-q_l^{'} \tag{1.9c}$$

at each discontinuity.

(ii) *Admissible traveling waves in the sense of Definition* 1.1 *are traveling waves in the sense of Definition* 1.2 *(or in the sense of Definition* 1.3*). If the admissible traveling wave is a piecewise smooth function with the first kind discontinuities, it satisfies not only the Rankine–Hugoniot condition* (1.9) *but also an entropy condition*

$$u_r<u_l \tag{1.10}$$

at each discontinuity.

(iii) *Suppose that the traveling wave* $(U,Q)(\xi)$ *in the sense of Definition* 1.2 *(or in the sense of Definition* 1.3*) is a piecewise smooth function with the first kind discontinuities and satisfies the entropy condition* (1.10) *at each discontinuity. Then* $(U,Q)(\xi)$ *is an admissible traveling wave in the sense of Definition* 1.1.

The proof of the previous proposition will be given in section 2. Here we note that (1.9b) implies the continuity of $Q(\xi)$. More precisely, it is shown in section 2 that $Q(\xi)$ is Lipschitz continuous even in the case where $U(\xi)$ has discontinuities. Also note that the entropy condition (1.10) is just the same as the one for the Burgers equation

$$u_t+uu_x=0.$$

By virtue of the entropy condition, when $u_+ < u_-$, we obtain the unique existence of a (discontinuous) traveling wave in the class of piecewise smooth functions whose discontinuities are of the first kind. This result can be stated as follows.

THEOREM 1.5.

(i) *Suppose that there exists an admissible traveling wave* $(U, Q)(x - st)$ *which satisfies the asymptotic condition* (1.6) *and is a piecewise smooth function with the first kind discontinuities. Then we have*

$$u_+ < u_-, \qquad s = \frac{1}{2}(u_+ + u_-) \tag{1.11}$$

and

$$\lim_{\xi \to \pm\infty} Q(\xi) = 0. \tag{1.12}$$

(ii) *Conversely, we suppose that* (1.11) *holds. Then there exists an admissible traveling wave* $(U, Q)(x - st)$ *satisfying* (1.6) *and* (1.12). *This admissible traveling wave is unique up to a shift in the class of piecewise smooth functions with the first kind discontinuities.*

(a) *When* $|u_+ - u_-| > \sqrt{2}$, $U(\xi)$ *is continuous except for one point, while* $Q(\xi)$ *is Lipschitz continuous. The conditions* (1.9), (1.10), *and* $u_r' = u_l' = -1$ *hold at the discontinuity of* $U(\xi)$.

(b) *If* $|u_+ - u_-| \le \sqrt{2}$, *then* $U(\xi)$ *is in* B^1 *and* $Q(\xi)$ *is in* B^2.

(c) *If* $|u_+ - u_-| < \frac{2\sqrt{2n}}{n+1}$, *then* $U(\xi)$ *is in* B^n *and* $Q(\xi)$ *is in* B^{n+1}, *where* $n = 2, 3, \ldots$.

Thus, when the shock strength $|u_+ - u_-|$ is less than or equal to $\sqrt{2}$, we have the C^1-smooth traveling waves. In proving the stability theorem in section 3, we require that $|u_+ - u_-| \le \frac{\sqrt{6}}{2}$ to make traveling waves C^3-smooth. In order to state the stability theorem, we define the function $(\phi, \psi)(x, t)$, expressing the perturbation from the traveling wave:

$$\phi(x, t) = u(x, t) - U(x - st) \tag{1.13a}$$

and

$$\psi(x, t) = q(x, t) - Q(x - st). \tag{1.13b}$$

As (u, q) and (U, Q) are solutions of (1.1), (ϕ, ψ) satisfies

$$\phi_t + \left(U\phi + \frac{1}{2}\phi^2\right)_x + \psi_x = 0, \tag{1.14a}$$

$$-\psi_{xx} + \psi + \phi_x = 0. \tag{1.14b}$$

We solve the above equations under the initial condition

$$\phi(x, 0) = \phi_0(x) \equiv u_0(x) - U(x), \tag{1.15}$$

where without loss of generality, we may assume that

$$\int_{-\infty}^{\infty} \phi_0(x)dx = 0. \tag{1.16}$$

In section 3, we prove the contraction property in L^1 that if $\phi_0 \in L^1$, then $\phi(\cdot,t) \in L^1$ and $|\phi(\cdot,t)|_1 \le |\phi_0|_1$ for each $t > 0$ (see Lemma 3.1). Based on this property, we define the integrated function Φ of ϕ as

$$\Phi(x,t) = \int_{-\infty}^{x} \phi(y,t)dy \tag{1.17}$$

and put

$$\Phi_0(x) = \int_{-\infty}^{x} \phi_0(y)dy. \tag{1.18}$$

The function Φ satisfies the following equation that is derived by integrating (1.14a) on the interval $(-\infty, x]$:

$$\Phi_t + U\Phi_x + \frac{1}{2}\Phi_x^2 + \psi = 0. \tag{1.19}$$

Now we state the stability result.

THEOREM 1.6. (i) *Let $u_+ < u_-$ and suppose that $|u_- - u_+| < \frac{\sqrt{6}}{2}$. Suppose also that $\phi_0 \in L^1 \cap H^2$ and $\Phi_0 \in L^2$. If $\|\Phi_0\|_3$ is sufficiently small, then the initial value problem* (1.14), (1.15) *has a unique global solution (ϕ, ψ) satisfying*

$$\lim_{t\to\infty} |(\phi,\psi)(\cdot,t)|_\infty = 0. \tag{1.20}$$

(ii) *In addition to the assumption of* (i)*, we suppose that $\Phi_0 \in L^1$. Then*

$$|(\phi,\psi)(\cdot,t)|_\infty = O(t^{-\frac{1}{4}}) \quad as \quad t \longrightarrow \infty. \tag{1.21}$$

The proof of the above theorem follows from the local existence theorem (Proposition 1.7) and a priori estimates (Propositions 1.8 and 1.9). As Proposition 1.7 is proved by the standard iteration method, we omit the proof. The derivation of the a priori estimates are given in section 3.

PROPOSITION 1.7. *Suppose that $U \in B^3$ and $\phi_0 \in H^2$. Then there exists a positive constant T_0, depending only on $\|\phi_0\|_2$, such that* (1.14), (1.15) *has a unique solution (ϕ, ψ) satisfying*

$$\phi \in C^0([0,T_0]; H^2) \cap C^1([0,T_0]; H^1) \tag{1.22a}$$

and

$$\psi \in C^0([0,T_0]; H^3). \tag{1.22b}$$

Furthermore, if $\phi_0 \in L^1$, then the solution verifies

$$\phi \in C^0([0,T_0]; L^1). \tag{1.22c}$$

By virtue of (1.22c), Φ is well defined by (1.17) and satisfies $\Phi \in C^0([0,T_0]; L^2)$ (and hence $\Phi \in C^0([0,T_0]; H^3)$), provided that $\Phi_0 \in L^2$. In this case our solution verifies

$$\|(\Phi,\psi)(t)\|_3 \le C\|\Phi_0\|_3 \quad \text{for} \quad t \in [0,T_0]. \tag{1.23}$$

We denote the supremum in time of the H^3 norm of Φ by N:

$$N(t) = \sup_{0\le\tau\le t} \|\Phi(\cdot,\tau)\|_3. \tag{1.24}$$

PROPOSITION 1.8. *Suppose that the assumptions of* (i) *in Theorem* 1.6 *holds. Let* $T > 0$ *and let* (ϕ,ψ) *be a solution of the problem* (1.14), (1.15), *which satisfies* (1.22) *with* T_0 *replaced by* T. *Then* $\Phi \in C^0([0,T];H^3)$. *Moreover, if* $N(T)$ *is so small that* $N(T) \le \frac{1}{10}$, *then we have the uniform estimate:*

$$\|(\Phi,\psi)(t)\|_3^2 + \int_0^t \|\phi(\tau)\|_2^2 + \|\psi(\tau)\|_3^2 d\tau + \int_0^t\int_{-\infty}^{\infty} |U_x|\Phi^2 dx\, d\tau \le C\|\Phi_0\|_3^2 \tag{1.25}$$

for $t \in [0,T]$, *where* C *is a positive constant independent of* T.

PROPOSITION 1.9. *Suppose that the assumptions of* (ii) *in Theorem* 1.6 *hold. Let* (ϕ,ψ) *be the global solution obtained in* (i) *of Theorem* 1.6. *Then we have the decay estimate:*

$$\|(\Phi,\psi)(t)\|_3 \le C(|\Phi_0|_1 + \|\Phi_0\|_3)(1+t)^{-1/4} \tag{1.26}$$

for $t \ge 0$.

Known results and outline of the paper. The system of equations (1.1) is derived as the third-order approximation of the full system describing the motion of radiating gas in thermo-nonequilibrium, while the second-order approximation gives the viscous Burgers equation $u_t + uu_x = u_{xx}$, and the first-order approximation gives the inviscid Burgers equation $u_t + uu_x = 0$. Hamer [3] studied these equations in the physical respect, especially for the steady progressive shock-wave solutions. Mathematically, Kawashima and Tanaka started the research of (1.1) in [6], which proved the local existence and uniqueness of the smooth solutions. Furthermore, under suitable conditions, [6] proved the global existence of smooth solutions and observed the asymptotic behaviors in the two cases $u_- = u_+$ and $u_- < u_+$. The first case gives diffusion waves which correspond to the viscous Burgers equation, while the second one gives rarefaction waves corresponding to the inviscid Burgers equation. On the other hand, Ito [4] proved the uniqueness and global existence of weak solutions in the space of functions of bounded variation. He also discussed the stability of rarefaction waves. These results in [6] and [4] indicate to us that the time asymptotic behavior for (1.1) is closely related to the one for the Burgers equation.

The purpose of this paper is to investigate the case $u_- > u_+$ for the same system. We will show the existence and stability of the traveling waves. It turns out that the condition $u_- > u_+$ is the necessary and sufficient condition to ensure the existence of traveling waves in weak sense. The magnitude of the quantity $|u_- - u_+|$ is shown to give information on the smoothness of the traveling waves. To obtain the smooth traveling waves, the condition $|u_- - u_+| \ll 1$ is required. The smaller $|u_- - u_+|$ gets, the smoother the traveling waves become. These phenomena are newly found by the authors. Although the discontinuous traveling waves are found and discussed in [8] and [10] for the relaxation models, these waves are C^∞ when they are differentiable. The phenomena mentioned above can occur because the viscosity derived from the elliptic equation (1.1b) is relatively weak, compared to that of the viscous conservation laws. However, when $|u_- - u_+|$ is small, the term q_{xx} in (1.1b) becomes small and our system (1.1) is well approximated by the viscous Burgers equation.

Furthermore, we prove the stability theorem which asserts that C^3-smooth traveling waves are time asymptotically stable, assuming that the antiderivative of the

initial perturbation is small in the Sobolev space H^3. Also, we obtain the convergence rate $t^{-1/4}$. The rate looks optimal, compared to the result given for the viscous Burgers equation.

The plan of this paper is as follows. After introducing the notations, we discuss the properties of traveling waves in section 2. In section 3, we give the proofs of the stability results. In section 2, we first discuss the unique existence of (discontinuous) traveling waves. The entropy condition (1.10), which is derived from Definition 1.1, play the essential role. Then we give necessary and sufficient conditions for making traveling waves differentiable. Furthermore, we show that the order of differentiability is determined by the shock strength $|u_- - u_+|$.

In section 3, we give the proof of stability theorem by proving the propositions in the case that $|u_- - u_+| < \frac{\sqrt{6}}{2}$ and the antiderivative of initial perturbation is small in the Sobolev space H^3. These results are given by applying the standard energy method not only to the system (1.14) but also to the integrated equation (1.19). The decay estimate is also given by the energy method which makes use of a time weight function.

Notations. For an integer $k \geq 0$, we denote by C^k the space of k-times continuously differentiable functions on $\mathbf{R}$. B^k denotes the space of C^k-functions whose derivatives up to order k are bounded. For $1 \leq p \leq \infty$, L^p denotes the usual Lebesgue space over $\mathbf{R}$ with the norm $|\cdot|_p$. For arbitrary integer $l \geq 0$, 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 $\|\cdot\|_0 = |\cdot|_2 \equiv \|\cdot\|$. We also denote by $C^k(I; H^l)$ the space of k-times continuously differentiable functions on the interval I with values in H^l.

We denote by $\mathcal{S}$ the space of all rapidly decreasing functions on $\mathbf{R}$. Also, ${C_0}^\infty$ denotes the space of C^∞-functions on $\mathbf{R}$, having compact supports.

Finally, by C or c we denote several constants without confusion.

2. Traveling waves. In this section we analyze the traveling waves. At first we discuss the properties of the traveling waves in weak sense to prove Proposition 1.4.

Traveling waves in weak sense (proof of Proposition 1.4). In this subsection we investigate definitions of traveling waves in weak sense. As we see function as distributions, we abbreviate that

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f g d\xi.$$

By direct calculation, it follows that the linear operator K in (1.3) is symmetric in the following sense.

LEMMA 2.1. *If $\psi \in \mathcal{S}$, then $K\psi \in \mathcal{S}$ and*

$$\langle KU, \psi \rangle = \langle U, K\psi \rangle$$

for arbitrary $U \in L^\infty$.

Proof of Proposition 1.4 (i). Assuming $(U, Q) \in L^\infty$ is a traveling wave in the sense of Definition 1.2, we have from (D.1b) that

$$\langle Q, \psi_\xi \rangle = \langle KU, \psi_{\xi\xi} \rangle = -\langle U - KU, \psi \rangle. \tag{2.1}$$

In deriving this equation, we used Lemma 2.1 and the equality $K\psi_{\xi\xi} = -(\psi - K\psi)$. By substituting (2.1) with $\psi = \phi \in C_0^\infty$ in (D.2), (D.3a) follows immediately.

By (2.1) and (D.1b),

$$\begin{aligned}\langle Q, -\psi_{\xi\xi} + \psi\rangle &= -\langle Q, \psi_{\xi\xi}\rangle + \langle Q, \psi\rangle \\ &= \langle U - KU, \psi_\xi\rangle + \langle KU, \psi_\xi\rangle \\ &= \langle U, \psi_\xi\rangle.\end{aligned} \tag{2.2}$$

This equation is (D.3b).

Next we assume that $(U, Q) \in L^\infty$ is a traveling wave in the sense of Definition 1.3. Let $\hat{\psi} = K\psi$ for arbitrary $\psi \in \mathcal{S}$. As $\psi = -\hat{\psi}_{\xi\xi} + \hat{\psi}$,

$$\begin{aligned}\langle Q, \psi\rangle &= \langle Q, -\hat{\psi}_{\xi\xi} + \hat{\psi}\rangle \\ &= \langle U, \hat{\psi}_\xi\rangle \\ &= \langle U, K\psi_\xi\rangle \\ &= \langle KU, \psi_\xi\rangle,\end{aligned} \tag{2.3}$$

where we used (D.3b) and Lemma 2.1. Thus we proved (D.1b). Noting that (D.1b) implies (2.1), we substitute (2.1) with $\psi = \phi \in {C_0}^\infty$ in (D.3a) to obtain (D.2). This proves the equivalence between Definitions 1.2 and 1.3.

The Rankine–Hugoniot condition (1.9) is proved by the standard computation employed in deriving the condition for the inviscid conservation laws. □

Let us note that $Q(\xi)$ for arbitrary traveling wave $(U, Q) \in L^\infty$ in the weak sense is not only continuous but also Lipschitz continuous because (2.1) implies $Q_\xi = U - KU \in L^\infty$.

Proof of Proposition 1.4 (ii) *and* (iii). As the first statement is apparent, we only prove the second statement of (ii) here. Let us assume that $(U, Q) \in L^\infty$ is an admissible traveling wave in the sense of Definition 1.1 and is a piecewise smooth function with the first kind discontinuities. Let $\{a_i; i = 1 \cdots N\}$ be the set of the discontinuous points of U contained in the support of $\varphi \in {C_0}^\infty$ with $\varphi \geq 0$. Also we denote by $u_{i,l}$ and $u_{i,r}$ the left and right limits of U at a_i, respectively. Since (U, Q) solves (D.2) and (D.1a), we see that (U, Q) satisfies the Rankine–Hugoniot condition (1.9) at each a_i and also the differential equations

$$-sU' + UU' + U - KU = 0 \tag{2.4a}$$

and

$$-Q'' + Q + U' = 0 \tag{2.4b}$$

on $\mathrm{supp}[\varphi] \setminus \{a_i\}$. Now the direct calculation using (1.9), (2.4a) shows that

$$\begin{aligned}\text{left-hand side of (D.1a)} = \frac{1}{2}\sum_i \{&sign(k - u_{i,r}) \\ &+ sign(u_{i,l} - k)\}\varphi(a_i)(k - u_{i,r})(u_{i,l} - k).\end{aligned} \tag{2.5}$$

In the above summation the suffix i runs over the set $\{i; u_{i,l} < k < u_{i,r}$ or $u_{i,l} > k > u_{i,r}\}$. Thus the entropy condition $u_{i,r} < u_{i,l}$ follows immediately from the observation using (2.5). The conclusion of (iii) follows from (i) and (2.5). The details are omitted. □

Existence of traveling waves (proof of Theorem 1.5). Here and after we look for the admissible traveling waves satisfying (1.6) in the class of piecewise smooth functions with the first kind discontinuities.

At first we suppose that there exists such an admissible traveling wave $(U,Q)(\xi)$. Integrating (1.7a) over the interval $[\xi,\xi']$, we have that

$$-sU(\xi)+\frac{1}{2}U(\xi)^2+Q(\xi)=-sU(\xi')+\frac{1}{2}U(\xi')^2+Q(\xi'). \tag{2.6}$$

Owing to the Rankine–Hugoniot conditions (1.9), the above equation holds true even if there are discontinuities in the interval $[\xi,\xi']$. Letting $\xi'\to\infty$ in (2.6) and using (1.6), we find that $Q(\xi')$ has a finite limit as $\xi'\to\infty$, which we denote by q_+. In a similar way, the limit $q_-=\lim_{\xi\to-\infty}Q(\xi)$ exists and is finite. Thus we get that

$$-sU(\xi)+\frac{1}{2}U(\xi)^2+Q(\xi)=-su_\pm+\frac{1}{2}u_\pm^2+q_\pm. \tag{2.7}$$

This equation implies that the state $(U,Q)(\xi)$ lies on the parabolic curve, which we denote by Γ, in the U-Q plane. Also we denote the apex of the curve Γ by $(\tilde{u},\tilde{q})$. Notice that $s=\tilde{u}$.

LEMMA 2.2. *If there exists a ξ_0 such that $(U,Q)(\xi)$ is differentiable for $\xi>\xi_0$ (resp., $\xi<\xi_0$), then $q_+=0$ (resp., $q_-=0$).*

Proof. We consider the case that $\xi>\xi_0$ only, since another case can be treated similarly. By (1.6),

$$u_+=\lim_{\xi\to\infty}U(\xi)=\lim_{\xi\to\infty}\frac{U(\xi)e^{-\xi}}{e^{-\xi}}=\lim_{\xi\to\infty}(U(\xi)-U'(\xi)).$$

Thus we have that $U'(\xi)\to 0$ as $\xi\to\infty$. Similarly, we see that $Q'(\xi)\to 0$ and $Q''(\xi)\to 0$ as $\xi\to\infty$. Substituting in (1.7b), we get the conclusion. ∎

Proof of Theorem 1.5 (i). It suffices to prove the following lemma.

LEMMA 2.3. (i) *$u_+<u_-$, $s=\frac{1}{2}(u_++u_-)$ and $q_\pm=0$.*

(ii) *Admissible traveling waves have at most one discontinuity.*

Proof. At first we prove that our admissible traveling waves have only finitely many discontinuities. Consider the case where $(u_+,q_+)\neq(\tilde{u},\tilde{q})$. We note that any admissible traveling wave $(U,Q)(\xi)$ laying on the curve Γ has the horizontal jump only in the U-Q plane, which goes from right to left, because it must satisfy the Rankine–Hugoniot condition (1.9b) and the entropy condition (1.10). On the other hand, the convergence (1.6) requires that $(U,Q)(\xi)$ is in a small neighborhood of (u_+,q_+) for any $\xi>R$, where R is a sufficiently large number. Since (u_+,q_+) is on the slope of Γ, we conclude that there is no discontinuity for $\xi>R$.

Similarly, if $(u_-,q_-)\neq(\tilde{u},\tilde{q})$, then there is no discontinuity for $\xi<-R$. Thus we have shown that when $(u_+,q_+)\neq(\tilde{u},\tilde{q})$ and $(u_-,q_-)\neq(\tilde{u},\tilde{q})$, the discontinuities appear only finitely many times. Consequently, applying Lemma 2.2, we have $q_\pm=0$ in this case.

Next we show that the case $(u_+,q_+)=(\tilde{u},\tilde{q})$ cannot occur by contradiction. Assume that $(u_+,q_+)=(\tilde{u},\tilde{q})$. Then $(u_-,q_-)\neq(\tilde{u},\tilde{q})$. It follows from the above consideration that there is no discontinuity for $\xi<-R$ so that $q_-=0$ by Lemma 2.2. Consequently, if we have infinitely many discontinuities, they should lie in a certain neighborhood of $(u_+,q_+)=(\tilde{u},\tilde{q})$. Since the discontinuity is a horizontal jump which goes from right to left in the U-Q plane, the point $(U,Q)(\xi)$ must move from left to right along the curve Γ, passing through the top $(\tilde{u},\tilde{q})$, as ξ increases from one jump point ξ_1 to the next one ξ_2. Therefore, there must exist a $\tilde{\xi}\in(\xi_1,\xi_2)$ such that $(U,Q)(\tilde{\xi})=(\tilde{u},\tilde{q})$. On the contrary, differentiating (1.7a) in ξ, we have that

$$Q''=(s-U)U''-(U')^2.$$

Substituting this in (1.7b) and evaluating the equation thus obtained at $\xi = \tilde{\xi}$, we find that

$$U'(\tilde{\xi}) = -U'(\tilde{\xi})^2 - \tilde{q} < 0,$$

where we used $\tilde{u} = s$ and $\tilde{q} > q_- = 0$.

This inequality implies that the point $(U, Q)(\xi)$ moves from right to left along the curve Γ around $(\tilde{u}, \tilde{q})$ as ξ increases. This is a contradiction. Therefore, we have only finitely many discontinuities and so $q_+ = 0$ by Lemma 2.2. This is a contradiction, too. Thus we have proved that the case $(u_+, q_+) = (\tilde{u}, \tilde{q})$ cannot occur. Similarly, we conclude that the case $(u_-, q_-) = (\tilde{u}, \tilde{q})$ does not happen, either.

Since $q_\pm = 0$, we have from (2.7) that $s = \frac{1}{2}(u_+ + u_-)$. Consequently, we see that $\tilde{q} = \frac{1}{8}|u_+ - u_-|^2 > 0$. By virtue of the property $\tilde{q} > 0$, we conclude that $u_+ < u_-$ by the same discussion. This proves (i). Also, employing the same discussion, we conclude that the jump can occur at most one point. The proof of the Lemma 2.3 is complete. □

Putting $s = \frac{1}{2}(u_+ + u_-)$ and $q_\pm = 0$ in (2.7), we can express Q in terms of U:

$$Q = -\frac{1}{2}(U - u_+)(U - u_-). \tag{2.8}$$

Substituting this in equation (1.7b), we have the second-order ordinary differential equation:

$$\left(U - \frac{u_+ + u_-}{2}\right) U'' + (U')^2 - \frac{1}{2}(U - u_+)(U - u_-) + U' = 0. \tag{2.9}$$

To simplify the equation, we change the dependent variable as $\hat{U} = U - \frac{u_+ + u_-}{2}$. Denoting this new variable $\hat{U}$ by U without confusion, we get the following system of the first-order ordinary differential equations:

$$\begin{pmatrix} U \\ V \end{pmatrix}' = \begin{pmatrix} V \\ -\frac{V^2 + V - \frac{1}{2}(U^2 - \alpha^2)}{U} \end{pmatrix}, \tag{2.10}$$

where $\alpha = \frac{|u_+ - u_-|}{2}$. Since $u_+ < u_-$, the asymptotic condition (1.6) is rewritten as

$$\lim_{\xi \to \pm\infty} (U, V)(\xi) = (\mp\alpha, 0). \tag{2.11}$$

We have to look for solutions of (2.10) under the condition (2.11). We denote by O the orbit of the solutions $(U, Q)(\xi)$ of (2.10), (2.11). Also we define the parabolic curve P in the U-V plane by

$$P = \left\{ (U, V); V^2 + V - \frac{1}{2}(U^2 - \alpha^2) = 0 \right\}. \tag{2.12}$$

By algebraic calculation, we know that the curve P intersects with the line $U = 0$ at the point $(0, v_0)$, where

$$v_0 = \frac{-1 + \sqrt{1 - 2\alpha^2}}{2}, \tag{2.13}$$

provided that $\alpha \le \frac{\sqrt{2}}{2}$, i.e., $|u_+ - u_-| \le \sqrt{2}$.

Otherwise, P does not meet the line $U = 0$.

The equilibrium points of (2.10) are $(\pm\alpha, 0)$. The Jacobian matrix of the vector on the right-hand side of (2.10), evaluated at $(\pm\alpha, 0)$, is

$$A_{\pm} = \begin{pmatrix} 0 & 1 \\ 1 & \mp\frac{1}{\alpha} \end{pmatrix}. \tag{2.14}$$

The eigenvalues of A_+ are

$$\mu_1 = \frac{-1+\sqrt{1+4\alpha^2}}{2\alpha} > 0, \quad \mu_2 = \frac{-1-\sqrt{1+4\alpha^2}}{2\alpha} < 0. \tag{2.15}$$

Thus $(\alpha, 0)$ is a saddle point. Similarly, another equilibrium point $(-\alpha, 0)$ is a saddle point, too.

LEMMA 2.4. *The trajectory O of solutions to* (2.10), (2.11) *is symmetric with respect to the line $U = 0$ in the U-V plane.*

Proof. Let $(U, V)(\xi)$ be a solution of (2.10) in a neighborhood of $\xi = \infty$ and satisfy the asymptotic condition (2.11) for $\xi \to \infty$. Then the function defined by

$$(\hat{U}, \hat{Q})(\xi) = (-U, Q)(-\xi)$$

is a solution of (2.10) around $\xi = -\infty$ and satisfies (2.11) for $\xi \to -\infty$. Since the equilibrium points $(\pm\alpha, 0)$ are saddle points, the conclusion follows easily. □

Lemma 2.4 and its proof show that if the shift in ξ is chosen appropriately, then U is an odd function and V is an even function with respect to ξ. Hence, it is enough to consider the case $U > 0$.

Next, comparing the eigenvector $(1, \mu_1)$ (row vector) of A_+ for μ_1 and the tangent of P, we have the following relation between the trajectory O and the curve P.

LEMMA 2.5. (i) *The trajectory O and the parabolic curve P do not intersect except for the points* $(\pm\alpha, 0)$ *and* $(0, v_0)$.

(ii) *The trajectory O is located above the curve P in the U-V plane.*

The above lemma implies that when $U > 0$, U and V are strictly decreasing functions of ξ.

Proof of Theorem 1.5 (ii-a). At first we prove that any admissible traveling wave satisfies

$$u'_r = u'_l = -1 \tag{2.16}$$

at each discontinuity. Let us consider the original dependent variable $(U, Q)(\xi)$. We have from (1.7a) that $q'_r = (s - u_r)u'_r$ and $q'_l = (s - u_l)u'_l$. Substituting these relations in (1.9c) and using $s = \frac{1}{2}(u_r + u_l)$, we see that

$$(u_l - u_r)\left(1 + \frac{1}{2}(u_l' + u_r')\right) = 0.$$

Since $u_l' = u_r'$ by Lemma 2.4, we have (2.16).

From now on we again write U in place of $U - \frac{1}{2}(u_+ + u_-)$. Consider the case where $|u_+ - u_-| > \sqrt{2}$. In this case the trajectory O goes to $-\infty$ as U goes to zero, and therefore any continuous traveling wave does not exist. But it is possible to connect the trajectory sprung from $(\alpha, 0)$ in $U > 0$ and the one from $(-\alpha, 0)$ in $U < 0$ by the horizontal jump between the two points (u_r, v_r) and (u_l, v_l). We require that

$$v_r = v_l = -1,$$

which ensures (2.16). As this jump satisfies the Rankine–Hugoniot condition (1.9) and the entropy condition (1.10) apparently, a discontinuous traveling wave exits. Thus we complete the proof of Theorem 1.5 (ii-a) with the fact that Q is Lipschitz continuous proved in the previous subsection. □

In order to obtain smooth traveling waves, we make a change of the independent variable as

$$U\frac{d}{d\xi} = \frac{d}{d\eta}$$

and transform (2.10) into

$$\frac{d}{d\eta}\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} UV \\ -V^2 - V + \frac{1}{2}(U^2 - \alpha^2) \end{pmatrix}. \tag{2.17}$$

This system has equilibrium points, $(\pm\alpha, 0)$ and $(0, v_0)$. The Jacobian matrix of the vector on the right-hand side of (2.17), evaluated at $(0, v_0)$, is given by

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}, \tag{2.18}$$

where

$$\lambda_1 = \frac{-1 + \sqrt{1 - 2\alpha^2}}{2} \equiv v_0 < 0 \quad \text{and} \quad \lambda_2 = -\sqrt{1 - 2\alpha^2} < 0. \tag{2.19}$$

The eigenvectors corresponding to λ_1 and λ_2 are

$$e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{2.20}$$

respectively.

Proof of Theorem 1.5 (ii-b). Let $|u_+ - u_-| \le \sqrt{2}$, i.e., $\alpha \le \frac{\sqrt{2}}{2}$. By the standard phase plane analysis, it is shown that the trajectory O connects the equilibrium points $(\alpha, 0)$ and $(0, v_0)$. Let us note that the change of the independent variable

$$\xi = -\int_\eta^\infty U(\zeta)d\zeta$$

gives the correspondence between $-\infty < \xi < 0$ and $-\infty < \eta < \infty$. Actually, this follows from the fact that $U(\eta) \to 0$ exponentially as $\eta \to \infty$. Obviously, $V \in B^0$ so that $U \in B^1$ as a function of ξ. Thus we have proved the existence of a desired traveling wave. □

Next, we show the following lemma.

LEMMA 2.6. (i) *If* $|u_+ - u_-| < \frac{4}{3}$, *then* $U \in B^2$.

(ii) *If* $|u_+ - u_-| < \frac{\sqrt{6}}{2}$, *then* $U \in B^3$.

Proof. By standard theory of autonomous ordinary differential equations, it is known that when $\lambda_1 < \lambda_2$, i.e., $|u_+ - u_-| > \frac{4}{3}$ (resp., $\lambda_1 > \lambda_2$, i.e., $|u_+ - u_-| < \frac{4}{3}$), the trajectory O tangents the eigenvector e_2 (resp., e_1) at the point $(0, v_0)$. Therefore, when $|u_+ - u_-| > \frac{4}{3}$, we see that $\frac{dV}{dU} \to \infty$ as $U \to 0_+$ so that $V \notin B^1$ and $U \notin B^2$ as functions of ξ. While in the case where $|u_+ - u_-| < \frac{4}{3}$, we have $\frac{dV}{dU} \to 0$ as $U \to 0_+$ so that $V \in B^1$ and $U \in B^2$. This proves (i).

To show (ii), we introduce

$$W_1 = \frac{V - \lambda_1}{U} \tag{2.21}$$

and transform (2.17) into

$$\frac{d}{d\eta}U = \lambda_1 U + U^2 W_1 \tag{2.22a}$$

and

$$\frac{d}{d\eta}W_1 = \frac{1}{2}U + (\lambda_2 - \lambda_1)W_1 - 2UW_1^2. \tag{2.22b}$$

We study the behavior of the trajectory of solutions for the system (2.22) around the equilibrium point $(0,0)$, which corresponds to $(0, v_0)$ for the system (2.17). The Jacobian matrix of the vector on the right-hand side of (2.22), evaluated at $(0,0)$, is

$$A_1 = \begin{pmatrix} \lambda_1 & 0 \\ \frac{1}{2} & \lambda_2 - \lambda_1 \end{pmatrix}. \tag{2.23}$$

The eigenvalues of A_1 are $\lambda_1^{(1)} = \lambda_1 < 0$ and $\lambda_2^{(1)} = \lambda_1 - \lambda_2 < 0$. The corresponding eigenvectors are

$$P_1^{(1)} = \begin{pmatrix} 1 \\ a_2 \end{pmatrix} \qquad \text{and} \qquad P_2^{(1)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = e_2, \tag{2.24}$$

respectively, where $a_2 = \frac{1}{2(2\lambda_1 - \lambda_2)}$. By the standard theory of autonomous ordinary differential equations, when $\lambda_1^{(1)} < \lambda_2^{(1)}$, i.e., $|u_+ - u_-| > \frac{\sqrt{6}}{2}$ (resp., $\lambda_1^{(1)} > \lambda_2^{(1)}$, i.e., $|u_+ - u_-| < \frac{\sqrt{6}}{2}$), the trajectory tangents the eigenvector $P_2^{(1)}$ (resp., $P_1^{(1)}$) at the point $(0,0)$. Therefore, if $|u_+ - u_-| > \frac{\sqrt{6}}{2}$, then $\frac{dW_1}{dU} \to \infty$ as $U \to 0_+$. This shows that $\frac{d^2V}{dU^2} \to \infty$ as $U \to 0_+$ so that $V \notin B^2$ and $U \notin B^3$ as functions of ξ. On the other hand, if $|u_+ - u_-| < \frac{\sqrt{6}}{2}$, then $\frac{dW_1}{dU} \to a_2$ as $U \to 0_+$. Consequently, we have $\frac{d^2V}{dU^2} \to 2a_2$ as $U \to 0_+$, which shows that $V \in B^2$ and $U \in B^3$. This completes the proof. □

Proof of Theorem 1.5 (ii-c). It suffices to prove the following theorem that gives a more precise statement of Theorem 1.5 (ii-c).

THEOREM 2.7. *Let $n \geq 1$ be an integer and suppose that*

$$|u_+ - u_-| < b_n \equiv \frac{2\sqrt{2(n+1)}}{n+2}. \tag{2.25}$$

Then V is a B^n-function of U and has the expansion

$$V = a_0 + a_1 U + a_2 U^2 + \cdots + a_n U^n + o(U^n) \tag{2.26}$$

for $U \to 0_+$, where the coefficients a_n are given by the formula (2.27) *below. Moreover, $V \in B^n$ and $U \in B^{n+1}$ as functions of ξ:*

$$a_0 = \lambda_1, \tag{2.27a}$$

$$a_2 = \frac{1}{2(2\lambda_1 - \lambda_2)}, \tag{2.27b}$$

and

$$a_n = \frac{n+2}{4(n\lambda_1 - \lambda_2)} \sum a_i a_{n-i} \quad \textit{for} \quad n = 4, 6, 8, \dots, \tag{2.27c}$$

where the summation is taken over all even integers i with $2 \le i \le n-2$, and

$$a_n = 0 \quad \textit{for} \quad n = 1, 3, 5, \dots. \tag{2.27d}$$

Proof. The theorem has already been proved for $n = 1$ and 2 in Lemma 2.6. We use the induction with respect to n to prove the general case. As the arguments are essentially the same as those in Lemma 2.6, we give only a brief sketch of the proof here.

Let $k \ge 1$ be an integer and assume that the theorem is valid for $n = k$. Then we prove the theorem for $n = k+1$. Let $|u_+ - u_-| < b_k$. Put

$$W_k = \frac{1}{U^k} \left(V - \sum_{j=0}^{k} a_j U^j \right). \tag{2.28}$$

Then $W_k \to 0$ as $U \to 0_+$ by the induction hypothesis. Write (2.28) as

$$V = \sum_{j=0}^{k} a_j U^j + U^k W_k, \tag{2.29}$$

and substitute this in (2.17). This yields the system of (U, W_k) of the form

$$\frac{d}{d\eta} \begin{pmatrix} U \\ W_k \end{pmatrix} = \begin{pmatrix} f_k(U, W_k) \\ g_k(U, W_k) \end{pmatrix}. \tag{2.30}$$

By direct computation, we see that $(0, 0)$ is an equilibrium point of (2.30). Moreover, the Jacobian matrix of the vector on the right-hand side of (2.30), evaluated at $(0, 0)$, is given by

$$A_k = \begin{pmatrix} \lambda_1 & 0 \\ s_k & \lambda_2 - k\lambda_1 \end{pmatrix},$$

where $s_k = 0$ if k is even and $s_k = \{(k+1)\lambda_1 - \lambda_2\} a_{k+1}$ if k is odd. The eigenvalues of A_k are $\lambda_1^{(k)} = \lambda_1$ and $\lambda_2^{(k)} = \lambda_2 - k\lambda_1$. The corresponding eigenvectors are

$$P_1^{(k)} = \begin{pmatrix} 1 \\ a_{k+1} \end{pmatrix} \qquad \text{and} \qquad P_2^{(k)} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{2.31}$$

respectively. Noting that $\lambda_1^{(k)} < \lambda_2^{(k)}$ (resp., $\lambda_1^{(k)} > \lambda_2^{(k)}$) is equivalent to $|u_+ - u_-| > b_{k+1}$ (resp., $|u_+ - u_-| < b_{k+1}$), we can deduce the desired conclusion of the theorem for $n = k+1$ by the same argument as in Lemma 2.6. In particular, we see that $U \in B^{k+2}$ for $|u_+ - u_-| < b_{k+1}$, but $U \notin B^{k+2}$ if $|u_+ - u_-| > b_{k+1}$. The proof is complete. □

Finally, in this section we summarize the result concerning the magnitude of traveling waves, which is necessary in the stability proof in section 3.

COROLLARY 2.8. *If* $|u_- - u_+| \le \sqrt{2}$, *then*

$$|U| \le \frac{1}{2}|u_- - u_+| \quad \textit{and} \quad |U'| \le \frac{1}{4}|u_- - u_+|^2. \tag{2.32a}$$

Moreover, we assume that $|u_- - u_+| < \frac{2\sqrt{2n}}{n+1}$ *for an integer* $n \ge 2$. *Then*

$$\left|\left(\frac{d}{d\xi}\right)^n U\right| \le C|u_- - u_+|^{n+1} \quad \textit{with} \quad C = \frac{C_0}{1-(n+1)|v_0|} \tag{2.32b}$$

and

$$|U'| \le |v_0| < \frac{1}{n+1}, \tag{2.33}$$

where C_0 *is a constant independent of* $|u_- - u_+|$ *and* v_0 *is the constant given in* (2.13) *with* $\alpha = \frac{|u_- - u_+|}{2}$.

Proof. It follows from Lemma 2.5 (ii) that $|U| \le \alpha$ and $|V| \le |v_0|$. Also, by simple calculation, using (2.13), we see that $|v_0| \le \alpha^2$ and that $|v_0| < \frac{1}{n+1}$ for $\alpha < \frac{\sqrt{2n}}{n+1}$. Thus we have proved (2.32a) and (2.33).

Next, we prove (2.32b) only for $n = 2$. The general case can be proved by the induction with respect to n. Observe that as U'' decays exponentially for $\xi \to \pm\infty$, $|U''|$ attains its maximum at a point ξ_0 where $U''' = V'' = 0$ holds. We have from (2.10) that

$$UV' + V^2 + V - \frac{1}{2}(U^2 - \alpha^2) = 0.$$

Differentiate the equation with respect to ξ and evaluate at $\xi = \xi_0$. Since $V''(\xi_0) = 0$, we obtain

$$U'' = V' = \frac{UV}{1+3V}$$

at $\xi = \xi_0$. Substituting the estimates $|U| \le \alpha$ and $|V| \le |v_0| \le \alpha^2$, we have

$$|U''| \le \frac{\alpha^3}{1-3|v_0|}$$

at $\xi = \xi_0$, which proves (2.32b) for $n = 2$. □

3. Stability. In this section, we discuss the stability of smooth traveling waves which are obtained in section 2. Here and after, we assume that $|u_- - u_+| < \frac{\sqrt{6}}{2}$ to make U of (1.6) being in B^3. Moreover, let (ϕ, ψ) be a solution of the problem (1.14), (1.15) satisfying (1.22) with T_0 replaced by T and assume that $N(T) \le \frac{1}{10}$.

***L*¹ estimates of *φ*.** At first, we give the L^1 estimate for ϕ. To this end, we introduce a_δ and A_δ by

$$a_\delta(\phi) = \rho_\delta * sign(\phi) \quad \text{and} \quad A_\delta(\phi) = \int_0^\phi a_\delta(\eta)d\eta, \tag{3.1}$$

where $\rho_\delta *$ denotes the Friedrichs mollifier. Note that $A_\delta(\phi) \to |\phi|$ as $\delta \to 0$. We have from (1.5b) that

$$\psi = -K\phi_x, \tag{3.2}$$

so that $\psi_x = \phi - K\phi$. Substitution of this in (1.14a) gives

$$\phi_t + \left(U\phi + \frac{1}{2}\phi^2\right)_x + \phi - K\phi = 0. \tag{3.3}$$

Multiplying $a_\delta(\phi)$ on this equation, we obtain that

$$\begin{aligned} &A_\delta(\phi)_t + U_x \int_0^\phi a'_\delta(\eta)\eta d\eta + a_\delta(\phi)(\phi - K\phi) \\ &\quad + \left\{ U\left(a_\delta(\phi)\phi - \int_0^\phi a'_\delta(\eta)\eta d\eta \right) + \int_0^\phi a_\delta(\eta)\eta d\eta \right\}_x = 0. \end{aligned} \tag{3.4}$$

We integrate this equation over $\mathbf{R} \times [0,t]$, obtaining

$$\begin{aligned} &\int_{-\infty}^{\infty} A_\delta(\phi)(x,t)dx + \int_0^t \int_{-\infty}^{\infty} U_x \left(\int_0^\phi a'_\delta(\eta)\eta d\eta \right) dx\, d\tau \\ &\qquad + \int_0^t \int_{-\infty}^{\infty} a_\delta(\phi)(\phi - K\phi) dx\, d\tau = \int_{-\infty}^{\infty} A_\delta(\phi_0) dx. \end{aligned} \tag{3.5}$$

Making $\delta \downarrow 0$, we have that

$$|\phi(t)|_1 + \int_0^t |\phi|_1 d\tau - \int_0^t \int_{-\infty}^{\infty} sign(\phi) K\phi dx\, d\tau = |\phi_0|_1. \tag{3.6}$$

Here we used the fact that second term on the left-hand side of (3.5) goes to zero as $\delta \to 0$ by the Lebesgue convergence theorem. Applying the estimate

$$\left| \int_{-\infty}^{\infty} sign(\phi) K\phi dx \right| \le |K\phi|_1 \le |\phi|_1$$

on (3.6), we obtain that $|\phi(t)|_1 \le |\phi_0|_1$. Thus we have proved the following lemma.

LEMMA 3.1. *If $\phi_0 \in L^1$, then*

$$|\phi(t)|_1 \le |\phi_0|_1. \tag{3.7}$$

***L^2* estimates (proof of Proposition 1.8).** Since we have assumed $\phi \in C^0([0,T]; L^1)$, Φ is well defined by (1.17) and satisfies $\Phi \in C^0([0,T]; L^2)$ if $\Phi_0 \in L^2$. Consequently, we have

$$\Phi \in C^0([0,T]; H^3), \tag{3.8}$$

and Φ satisfies (1.19). First we apply the standard energy method to equation (1.19) to derive the L^2-estimate for Φ. Multiplying Φ on (1.19), we have

$$\left(\frac{1}{2}\Phi^2\right)_t - \frac{1}{2}U_x\Phi^2 + \frac{1}{2}\Phi\phi^2 + \Phi\psi + \left(\frac{1}{2}U\Phi^2\right)_x = 0. \tag{3.9}$$

Since we have from (1.14b) that

$$\Phi\psi = \phi^2 - (\psi^2 + \psi_x^2) + \{(\Phi + \psi)(\psi_x - \phi)\}_x,$$

(3.9) is modified as

$$\left(\frac{1}{2}\Phi^2\right)_t - \frac{1}{2}U_x\Phi^2 + \phi^2 - (\psi^2 + \psi_x^2) + \frac{1}{2}\Phi\phi^2 + (\cdots)_x = 0, \tag{3.10}$$

where here and in what follows, $(\cdots)_x$ denotes terms which vanish after integration over $x \in \mathbf{R}$. We integrate (3.10) over $\mathbf{R} \times [0, t]$, obtaining

(3.11)
$$|\Phi(t)|_2^2 + 2\int_0^t |\phi|_2^2 d\tau + \int_0^t\int_{-\infty}^{\infty} |U_x|\Phi^2 dx\, d\tau = |\Phi_0|_2^2 + 2\int_0^t \|\psi\|_1^2 d\tau - \int_0^t\int_{-\infty}^{\infty} \Phi\phi^2 dx\, d\tau,$$

where we used the fact that $U_x < 0$, namely, U is a strictly decreasing function of $\xi = x - st$.

Next we derive the L^2-estimate for ϕ in a similar way. Multiplying ϕ and ψ on equations (1.14a) and (1.14b), respectively, and adding these two equations, we have

$$\left(\frac{1}{2}\phi^2\right)_t + \psi^2 + \psi_x^2 + \frac{1}{2}U_x\phi^2 + \left\{\frac{1}{2}U\phi^2 + \frac{1}{3}\phi^3 - \psi(\psi_x - \phi)\right\}_x = 0. \tag{3.12}$$

Integrating (3.12) over $\mathbf{R} \times [0, t]$ and using the fact that $U_x < 0$, we obtain

$$|\phi(t)|_2^2 + 2\int_0^t \|\psi\|_1^2 d\tau = |\phi_0|_2^2 + \int_0^t\int_{-\infty}^{\infty} |U_x|\phi^2 dx\, d\tau. \tag{3.13}$$

Now we add (3.11) and (3.13), obtaining

$$\begin{aligned} |\Phi(t)|_2^2 + |\phi(t)|_2^2 + 2\int_0^t |\phi|_2^2 d\tau + \int_0^t\int_{-\infty}^{\infty} |U_x|\Phi^2 dx\, d\tau &= |\Phi_0|_2^2 + |\phi_0|_2^2 \\ &\quad + \int_0^t\int_{-\infty}^{\infty} (|U_x| - \Phi)\phi^2 dx\, d\tau. \end{aligned} \tag{3.14}$$

We would like to estimate the last term on the right-hand side of (3.14). We recall that when $|u_+ - u_-| < \frac{\sqrt{6}}{2}$, Corollary 2.8 gives that

$$|U_x| < \frac{1}{4}, \qquad |U_{xx}| \le C, \tag{3.15a}$$

and

$$|U_{xxx}| < CM \quad \text{with} \quad M = \frac{1}{1 - 4|v_0|}. \tag{3.15b}$$

Using (3.15a) and $N(T) \le \frac{1}{10}$, we see that $|U_x| + |\Phi| < 1$. Therefore, we have from (3.14) that

$$|\Phi(t)|_2^2 + |\phi(t)|_2^2 + \int_0^t |\phi|_2^2 d\tau + \int_0^t\int_{-\infty}^{\infty} |U_x|\Phi^2 dx\, d\tau \le |\Phi_0|_2^2 + |\phi_0|_2^2. \tag{3.16}$$

Moreover, applying this estimate to the right-hand side of (3.13), we obtain

$$\int_0^t \|\psi\|_1^2 d\tau \le |\Phi_0|_2^2 + |\phi_0|_2^2. \tag{3.17}$$

Next we derive the estimate for ϕ_x. We differentiate (1.14) with respect to x and multiply ϕ_x and ψ_x on the first and the second equations thus obtained, respectively. Then we add these two equations, getting

$$\begin{aligned}\left(\frac{1}{2}\phi_x^2\right)_t &+ \psi_x^2 + \psi_{xx}^2 + \frac{3}{2}U_x\phi_x^2 + U_{xx}\phi\phi_x + \frac{1}{2}\phi_x^3 \\ &+ \left\{\frac{1}{2}U\phi_x^2 + \frac{1}{2}\phi\phi_x^2 - \psi_x(\psi_{xx} - \phi_x)\right\}_x = 0.\end{aligned} \tag{3.18}$$

On the other hand, by squaring (1.14b), we have

$$\phi_x^2 = \psi^2 + 2\psi_x^2 + \psi_{xx}^2 - 2(\psi\psi_x)_x. \tag{3.19}$$

We rewrite (3.18) by using (3.19) as

$$\left(\frac{1}{2}\phi_x^2\right)_t + \phi_x^2 - (\psi^2 + \psi_x^2) + \frac{3}{2}U_x\phi_x^2 + U_{xx}\phi\phi_x + \frac{1}{2}\phi_x^3 + (\cdots)_x = 0. \tag{3.20}$$

Here, applying (3.15a) and $N(T) \le \frac{1}{10}$, we see that

$$\left|\frac{3}{2}U_x\phi_x^2 + U_{xx}\phi\phi_x + \frac{1}{2}\phi_x^3\right| \le \frac{1}{2}\phi_x^2 + C\phi^2. \tag{3.21}$$

We integrate (3.20) over $\mathbf{R}\times[0,t]$ and substitute (3.21) in the resulting equation. This yields

$$|\phi_x(t)|_2^2 + \int_0^t |\phi_x|_2^2 d\tau \le |\phi_0'|_2^2 + C\int_0^t |\phi(\tau)|_2^2 d\tau + 2\int_0^t \|\psi(\tau)\|_1^2 d\tau \le C(|\Phi_0|_2^2 + \|\phi_0\|_1^2), \tag{3.22}$$

where we used the estimate (3.16) and (3.17). Also, from (3.19) and (3.22) we have

$$\int_0^t \|\psi(\tau)\|_2^2 d\tau \le C(|\Phi_0|_2^2 + \|\phi_0\|_1^2). \tag{3.23}$$

Finally, we derive the L^2-estimate for ϕ_{xx} by the similar method used in deriving the estimate for ϕ_x. We differentiate (1.14) twice with respect to x and multiply ϕ_{xx} and ψ_{xx} on the first and the second equations thus obtained, respectively. Then, adding the resulting two equations, we have

$$\begin{aligned}\left(\frac{1}{2}\phi_{xx}^2\right)_t &+ \psi_{xx}^2 + \psi_{xxx}^2 + \frac{5}{2}U_x\phi_{xx}^2 + 3U_{xx}\phi_x\phi_{xx} \\ &+ U_{xxx}\phi\phi_{xx}^2 + \frac{5}{2}\phi_x\phi_{xx}^2 + (\dots)_x = 0.\end{aligned} \tag{3.24}$$

Also, similarly to (3.19), we have

$$\phi_{xx}^2 = \psi_x^2 + 2\psi_{xx}^2 + \psi_{xxx}^2 - 2(\psi_x\psi_{xx})_x. \tag{3.25}$$

We rewrite (3.24) by using (3.25) and obtain as a counterpart of (3.20)

$$\begin{aligned}\left(\frac{1}{2}\phi_{xx}^2\right)_t + \phi_{xx}^2 - (\psi_x^2+\psi_{xx}^2) + \frac{5}{2}U_x\phi_{xx}^2 + 3U_{xx}\phi_x\phi_{xx} \\ + U_{xxx}\phi\phi_{xx}^2 + \frac{5}{2}\phi_x\phi_{xx}^2 + (\dots)_x = 0.\end{aligned} \tag{3.26}$$

Here, using (3.15) and $N(T) \le \frac{1}{10}$, we see that

$$\left|\frac{5}{2}U_x\phi_{xx}^2 + 3U_{xx}\phi_x\phi_{xx} + U_{xxx}\phi\phi_{xx} + \frac{5}{2}\phi_x\phi_{xx}^2\right| \le \frac{15}{16}\phi_{xx}^2 + CM^2\phi^2 + C\phi_x^2. \tag{3.27}$$

Now, integrating (3.26) over $\mathbf{R}\times[0,t]$ and substituting (3.27) in the resulting equation, we obtain

$$\begin{aligned}|\phi_{xx}(t)|_2^2 + \int_0^t |\phi_{xx}(\tau)|_2^2 d\tau &\le C|\phi_0''|_2^2 + C(M)\int_0^t \|\phi(\tau)\|_1^2 + \|\psi_x(\tau)\|_1^2 d\tau \\ &\le C(M)(|\Phi_0|_2^2 + \|\phi_0\|_2^2),\end{aligned} \tag{3.28}$$

where we used the estimates (3.16), (3.22), and (3.23). Here $C(M)$ denotes a constant depending on M in (3.15b). Also, from (3.25) and (3.28) we have

$$\int_0^t \|\psi_x(\tau)\|_2^2 \le C(M)(|\Phi_0|_2^2 + \|\phi_0\|_2^2). \tag{3.29}$$

To complete the proof of Proposition 1.8, we combine (3.16), (3.22), (3.23), (3.28), and (3.29), obtaining

$$\begin{aligned}|\Phi(t)|_2^2 + \|\phi(t)\|_2^2 + \int_0^t \|\phi(\tau)\|_2^2 + \|\psi(\tau)\|_3^2 d\tau \\ + \int_0^t\int_{-\infty}^{\infty} |U_x|\Phi^2 dx\, d\tau \le C(M)(|\Phi_0|_2^2 + \|\phi_0\|_2^2).\end{aligned} \tag{3.30}$$

This inequality combined with the inequality $\|\psi\|_3 \le \|\phi_x\|_1$, which is derived from (3.19) and (3.25), gives the desired estimate (1.25). This completes the proof of Proposition 1.8 and hence of Theorem 1.6 (i).

Decay estimates (proof of Proposition 1.9). In this subsection we derive the decay estimate (1.26). Let (ϕ,ψ) be the global solution to the problem (1.14), (1.15), which is obtained in Theorem 1.6. Note that this solution verifies

$$\sup_{0\le t\le\infty} \|\Phi(t)\|_3 \le C\|\Phi_0\|_3 \le \frac{1}{10}. \tag{3.31}$$

At first we prove the L^1-estimate of Φ, which plays an important role in getting the convergence rate.

LEMMA 3.2. *If $\Phi_0 \in L^1$, then $\Phi \in C^0([0,\infty); L^1)$ and*

$$|\Phi(t)|_1 + \int_0^t\int_{-\infty}^{\infty} |U_x||\Phi| dx\, d\tau \le |\Phi_0|_1 + \frac{1}{2}(|\Phi_0|_2^2 + |\phi_0|_2^2). \tag{3.32}$$

Proof. From (3.2), we have that

$$\psi_x = \phi - K\phi = (\Phi - K\Phi)_x.$$

By integrating this equation on the interval $(-\infty, x]$, we obtain

$$\psi = \Phi - K\Phi. \tag{3.33}$$

Substituting this equation in (1.19), we have that

$$\Phi_t + U\Phi_x + \frac{1}{2}\phi^2 + \Phi - K\Phi = 0. \tag{3.34}$$

Multiplying $a_\delta(\Phi)$ on this equation and integrating over $\mathbf{R} \times [0,t]$, we obtain that

$$\begin{aligned} &\int_{-\infty}^{\infty} A_\delta(\Phi(t))dx + \int_0^t \int_{-\infty}^{\infty} |U_x| A_\delta(\Phi) dx\, d\tau + \frac{1}{2}\int_0^t \int_{-\infty}^{\infty} a_\delta(\Phi)\phi^2 dx\, d\tau \\ &\qquad + \int_0^t \int_{-\infty}^{\infty} a_\delta(\Phi)(\Phi - K\Phi) dx\, d\tau = \int_{-\infty}^{\infty} A_\delta(\phi_0) dx. \end{aligned} \tag{3.35}$$

Letting $\delta \downarrow 0$ in (3.35), we have

$$|\Phi(t)|_1 + \int_0^t \int_{-\infty}^{\infty} |U_x||\Phi| dx\, d\tau \le |\Phi_0|_1 + \frac{1}{2}\int_0^t |\phi|_2^2 d\tau \le |\Phi_0|_1 + \frac{1}{2}(|\Phi_0|_2^2 + |\phi_0|_2^2),$$

where we used the estimate (3.16). This completes the proof. □

Next we derive the decay estimate for Φ and ϕ. Let $\mu > \frac{1}{2}$ be a sufficiently large number which is fixed. We multiply $(\tau+1)^\mu$ on equations (3.10) and (3.12), and integrate the resulting equations over $\mathbf{R} \times [0,t]$. This yields

$$\begin{aligned} &(t+1)^\mu |\Phi(t)|_2^2 + 2\int_0^t (\tau+1)^\mu |\phi(\tau)|_2^2 d\tau \\ &\qquad + \int_0^t \int_{-\infty}^{\infty} (\tau+1)^\mu |U_x| \Phi^2 dx\, d\tau = |\Phi_0|_2^2 + \mu \int_0^t (\tau+1)^{\mu-1} |\Phi(\tau)|_2^2 d\tau \\ &\qquad + 2\int_0^t (\tau+1)^\mu \|\psi(\tau)\|_1^2 d\tau - \int_0^t \int_{-\infty}^{\infty} (\tau+1)^\mu \Phi\phi^2 dx\, d\tau. \end{aligned} \tag{3.36}$$

and

$$\begin{aligned} &(t+1)^\mu |\phi(t)|_2^2 + 2\int_0^t (\tau+1)^\mu \|\psi(\tau)\|_1^2 d\tau = |\phi_0|_2^2 + \mu \int_0^t (\tau+1)^{\mu-1} |\phi(\tau)|_2^2 d\tau \\ &\qquad + \int_0^t \int_{-\infty}^{\infty} (\tau+1)^\mu |U_x| \phi^2 dx\, d\tau. \end{aligned} \tag{3.37}$$

Adding the above two equations, we obtain

$$\begin{aligned} &(t+1)^\mu (|\Phi(t)|_2^2 + |\phi(t)|_2^2) + 2\int_0^t (\tau+1)^\mu |\phi(\tau)|_2^2 d\tau \\ &\qquad + \int_0^t \int_{-\infty}^{\infty} (\tau+1)^\mu |U_x| \Phi^2 dx\, d\tau = |\Phi_0|_2^2 + |\phi_0|_2^2 \\ &\qquad + \mu \int_0^t (\tau+1)^{\mu-1} (|\Phi(\tau)|_2^2 + |\phi(\tau)|_2^2) d\tau \\ &\qquad + \int_0^t \int_{-\infty}^{\infty} (\tau+1)^\mu (|U_x| - \Phi)\phi^2 dx\, d\tau. \end{aligned} \tag{3.38}$$

We estimate the integrals on the right-hand side of (3.38). First, using (3.15a) and (3.31), we have

$$\int_0^t\int_{-\infty}^{\infty}(\tau+1)^{\mu}(|U_x|-\Phi)\phi^2 dx\,d\tau \le \frac{1}{2}\int_0^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau. \tag{3.39}$$

Next, applying the Gagliardo–Nirenberg inequality

$$|\Phi|_2 \le C|\Phi_x|_2^{\frac{1}{3}}|\Phi|_1^{\frac{2}{3}} = C|\phi|_2^{\frac{1}{3}}|\Phi|_1^{\frac{2}{3}}$$

and the Hölder inequality, we see that

$$\begin{aligned}\mu\int_0^t(\tau+1)^{\mu-1}|\Phi(\tau)|_2^2 d\tau &\le C\sup_{\tau}|\Phi(\tau)|_1^{\frac{4}{3}}\int_0^t(\tau+1)^{\mu-1}|\phi(\tau)|_2^{\frac{2}{3}}d\tau\\ &\le C\sup_{\tau}|\Phi(\tau)|_1^{\frac{4}{3}}\left[\int_0^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau\right]^{\frac{1}{3}}\left[\int_0^t(\tau+1)^{\mu-\frac{2}{3}}d\tau\right]^{\frac{2}{3}}\\ &\le \varepsilon\int_0^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau + C(\varepsilon)(t+1)^{\mu-\frac{1}{2}}\sup_{0\le\tau\le t}|\Phi(\tau)|_1^2 \end{aligned}\tag{3.40}$$

for arbitrary $\varepsilon>0$, where $C(\varepsilon)$ is a constant depending on ε. Here we assumed that $\mu>\frac{1}{2}$. Finally, we show that

$$\mu\int_0^t(\tau+1)^{\mu-1}|\phi(\tau)|_2^2 d\tau \le \varepsilon\int_0^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau + C(\varepsilon)\int_0^t|\phi(\tau)|_2^2 d\tau \tag{3.41}$$

for arbitrary $\varepsilon>0$. To see this, we choose T such that $\mu(T+1)^{-1}=\varepsilon$ and divide the integral on the left-hand side of (3.41) into two parts corresponding to $[0,T]$ and $[T,t]$, respectively. The first part is estimated as

$$\mu\int_0^T(\tau+1)^{\mu-1}|\phi(\tau)|_2^2 d\tau \le \mu(T+1)^{\mu-1}\int_0^T|\phi(\tau)|_2^2 d\tau \le C(\varepsilon)\int_0^t|\phi(\tau)|_2^2 d\tau,$$

while the second is

$$\mu\int_T^t(\tau+1)^{\mu-1}|\phi(\tau)|_2^2 d\tau \le \mu(T+1)^{-1}\int_T^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau \le \varepsilon\int_0^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau.$$

This proves (3.41). Now, substitute (3.39), (3.40), and (3.41) in (3.38). Then letting $\varepsilon=\frac{1}{4}$ on the inequality thus obtained, we obtain

$$\begin{aligned}&(t+1)^{\mu}(|\Phi(t)|_2^2+|\phi(t)|_2^2)+\int_0^t(\tau+1)^{\mu}|\phi(\tau)|_2^2 d\tau+\int_0^t\int_{-\infty}^{\infty}(\tau+1)^{\mu}|U_x|\Phi^2 dx\,d\tau\\ &\qquad\le |\Phi_0|_2^2+|\phi_0|_2^2+C(t+1)^{\mu-\frac{1}{2}}\sup_{0\le\tau\le t}|\Phi(\tau)|_1^2+C\int_0^t|\phi(\tau)|_2^2 d\tau\\ &\qquad\le C(t+1)^{\mu-\frac{1}{2}}(|\Phi_0|_1^2+|\Phi_0|_2^2+|\phi_0|_2^2),\end{aligned}\tag{3.42}$$

where we used the estimates (3.32) and (3.16). Moreover, applying (3.42) to the right-hand side of (3.37), we have

$$\int_0^t(\tau+1)^{\mu}\|\psi(\tau)\|_1^2 d\tau \le C(t+1)^{\mu-\frac{1}{2}}(|\Phi_0|_1^2+|\Phi_0|_2^2+|\phi_0|_2^2). \tag{3.43}$$

Next we derive the decay estimate for ϕ_x by multiplying $(\tau+1)^\mu$ on (3.20) and integrating it over $\mathbf{R}\times[0,t]$. Then, using (3.21), we obtain

$$(t+1)^\mu|\phi_x(t)|_2^2+\int_0^t(\tau+1)^\mu|\phi_x(\tau)|_2^2d\tau\le|\phi_0^{'}|_2^2+\mu\int_0^t(\tau+1)^{\mu-1}|\phi_x(\tau)|_2^2d\tau$$
$$+C\int_0^t(\tau+1)^\mu|\phi_x(\tau)|_2^2d\tau+2\int_0^t(\tau+1)^\mu\|\psi(\tau)\|_1^2d\tau. \tag{3.44}$$

We substitute the estimate (3.41) with ϕ replaced by ϕ_x to the second term on the right-hand side of (3.44). Then, using (3.22), (3.42), and (3.43), we obtain

$$(t+1)^\mu|\phi_x(t)|_2^2+\int_0^t(\tau+1)^\mu|\phi_x(\tau)|_2^2d\tau\le C(t+1)^{\mu-\frac12}(|\Phi_0|_1^2+|\Phi_0|_2^2+\|\phi_0\|_1^2). \tag{3.45}$$

Also, from (3.19) and (3.45) we have

$$\int_0^t(\tau+1)^\mu\|\psi(\tau)\|_2^2d\tau\le C(t+1)^{\mu-\frac12}(|\Phi_0|_1^2+|\Phi_0|_2^2+\|\phi_0\|_1^2). \tag{3.46}$$

Similarly, we derive the decay for ϕ_{xx}. We multiply $(\tau+1)^\mu$ on equation (3.26) and integrate it over $\mathbf{R}\times[0,t]$. Then, using (3.27), we obtain

$$(t+1)^\mu|\phi_{xx}(t)|_2^2+\int_0^t(\tau+1)^\mu|\phi_{xx}(\tau)|_2^2d\tau\le C|\phi_0^{''}|_2^2+\mu C\int_0^t(\tau+1)^{\mu-1}|\phi_{xx}(\tau)|_2^2d\tau$$
$$+C(M)\int_0^t(\tau+1)^\mu(\|\phi(\tau)\|_1^2+\|\psi_x(\tau)\|_1^2)d\tau. \tag{3.47}$$

By applying (3.41) with ϕ replaced by ϕ_{xx} and using (3.28), (3.42), (3.45), and (3.46), we have from (3.47) that

$$(t+1)^\mu|\phi_{xx}(t)|_2^2+\int_0^t(\tau+1)^\mu|\phi_{xx}(\tau)|_2^2d\tau\le C(M)(t+1)^{\mu-\frac12}(|\Phi_0|_1^2+|\Phi_0|_2^2+\|\phi_0\|_2^2). \tag{3.48}$$

Also, from (3.25) and (3.48) we have

$$\int_0^t(\tau+1)^\mu\|\psi_x(\tau)\|_2^2d\tau\le C(M)(t+1)^{\mu-\frac12}(|\Phi_0|_1^2+|\Phi_0|_2^2+\|\phi_0\|_2^2). \tag{3.49}$$

Finally, combining (3.42), (3.45), (3.46), (3.48), and (3.49), we arrive at the estimate

$$\begin{aligned}(t+1)^\mu(|\Phi(t)|_2^2+\|\phi(t)\|_2^2)+\int_0^t(\tau+1)^\mu(\|\phi(\tau)\|_2^2+\|\psi(\tau)\|_3^2)d\tau\\ +\int_0^t\int_{-\infty}^\infty(\tau+1)^\mu|U_x|\Phi^2dx\,d\tau\\ \le C(M)(t+1)^{\mu-\frac12}(|\Phi_0|_1^2+|\Phi_0|_2^2+\|\phi_0\|_2^2).\end{aligned} \tag{3.50}$$

In particular, we have shown that

$$|\Phi(t)|_2+\|\phi(t)\|_2\le C(M)(t+1)^{-\frac14}(|\Phi_0|_1+|\Phi_0|_2+\|\phi_0\|_2), \tag{3.51}$$

which together with the inequality $\|\psi\|_3\le C\|\phi\|_2$ gives

$$\|\psi(t)\|_3\le C(M)(t+1)^{-\frac14}(|\Phi_0|_1+|\Phi_0|_2+\|\phi_0\|_2). \tag{3.52}$$

This completes the proof of Proposition 1.9 and hence of Theorem 1.6 (ii).

Acknowledgment. The authors would like to express their deepest gratitude to Professor Xu-Yan Chen for stimulating discussions.

REFERENCES

[1] A. T. Cates and D. G. Crighton, *Nonlinear diffraction and caustic formation*, in Proc. Roy. Soc. London Ser. A, 430 (1990), pp. 69–88.
[2] D. G. Crighton, *Model equations of nonlinear acoustics*, Ann. Rev. Fluid Mech., 11 (1979), pp. 11–33.
[3] K. Hamer, *Nonlinear effects on the propagation of sound waves in a radiating gas*, Quart. J. Mech. Appl. Math., 24 (1971), pp. 155–168.
[4] K. Ito, *BV-solutions of the hyperbolic-elliptic system for a radiating gas*, to appear.
[5] S. Kawashima, *Large-time behavior of solutions to hyperbolic-parabolic system of conservation laws and applications*, in Proc. Roy. Soc. Edinburgh, 106A (1987), pp. 169–194.
[6] S. Kawashima and Y. Tanaka, *Asymptotic behavior of solutions to the one-dimensional model system for radiating gas*, unpublished note.
[7] S. N. Kruzkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.
[8] T.-P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.
[9] A. Matsumura and K. Nishihara, *Asymptotic stability of traveling waves for scalar viscous conservation laws with non-convex nonlinearity*, Comm. Math. Phys., 165 (1994), pp. 83–96.
[10] S. Nishibata, *The initial boundary value problems for hyperbolic conservation laws with relaxation*, J. Differential Equations, 130 (1996), pp. 100–126.
[11] W. G. Vincenti and C. H. Kruger, *Introduction to Physical Gas Dynamics*, Wiley, New York, 1965.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 118–139

ENCAPSULATED-VORTEX SOLUTIONS TO EQUIVARIANT WAVE EQUATIONS: EXISTENCE*

JOSEPH IAIA† AND HENRY WARCHALL†

Abstract. We prove the existence of vector-valued exponentially localized solutions to a class of nonlinear elliptic equations with equivariant nonlinearities. The elliptic equations govern the spatial profiles of solitary wave solutions to nonlinear wave equations with globally regular solutions. These solitary waves are localized versions of nonlinear Schrödinger vortices. We use constructive methods to show that there are localized solutions with any prescribed number of nodal surfaces.

Key words. nonlinear wave equations, solitary waves, multidimensional solitons, vortices, equivariant wave equations, localized traveling waves

AMS subject classifications. 34B15, 35L70, 35Q53, 35Q55

PII. S0036141097316925

1. Introduction.

We consider nonlinear Schrödinger (NLS) and nonlinear Klein–Gordon (NLKG) equations

$$-J\,u_t - \Delta u = \vec{g}(u) \qquad \text{(NLS)} \tag{1.1}$$

and

$$u_{tt} - \Delta u = \vec{g}(u) \qquad \text{(NLKG)}, \tag{1.2}$$

where J is an invertible real skew-symmetric $M \times M$ matrix, and solutions u are $\mathbb{R}^M$-valued functions on spacetime $\mathbb{R}^{N+1}$. The nonlinearity $\vec{g} : \mathbb{R}^M \to \mathbb{R}^M$ is assumed to be a continuous radial vector field; in particular, $\vec{g}(0) = 0$ and

$$\vec{g}(y) = g(|y|)\hat{y} \qquad \text{for} \quad y \neq 0, \tag{1.3}$$

where $g : [0, \infty) \to \mathbb{R}$ is a continuous function with $g(0) = 0$ and $\hat{y} \equiv \frac{y}{|y|}$.

We study standing-wave solutions

$$u(x,t) \equiv e^{\nu t K}\, v(x) \tag{1.4}$$

to the nonlinear wave equations (NLS) and (NLKG), where ν is a real constant, K is a real skew-symmetric $M \times M$ matrix, and $v : \mathbb{R}^N \to \mathbb{R}^M$. For (NLS), we take $K \equiv J^{-1}$, so that $-J\frac{\partial}{\partial t}u = -\nu\, u$. For (NLKG), we take K to be skew-symmetric and such that $K^2 = -I$, so that $\frac{\partial^2}{\partial t^2}u = -\nu^2 u$. (These requirements on J and K necessitate that the range space $\mathbb{R}^M$ be even-dimensional.)

Because $e^{\nu t K}$ is an isometry on the Euclidean space $\mathbb{R}^M$, it follows that $\vec{g}\left(e^{\nu t K}\, v(x)\right) = e^{\nu t K}\vec{g}\,(v(x))$, and consequently, if $u(x,t) \equiv e^{\nu t K}\, v(x)$ is a solution of (NLS) or (NLKG), then the function v must satisfy the elliptic equation

$$\Delta v + \vec{f}(v) = 0, \tag{1.5}$$

*Received by the editors February 19, 1997; accepted for publication (in revised form) February 9, 1998; published electronically October 14, 1998. This work was partially supported by University of North Texas Faculty Research Grants.

http://www.siam.org/journals/sima/30-1/31692.html

†Department of Mathematics, University of North Texas, Denton, TX 76203-5118 (iaia@cas.unt.edu, warchall@unt.edu).

where $\vec{f}(y) \equiv \vec{g}(y) + \omega\, y$, with

$$\omega \equiv \begin{cases} \nu & \text{for NLS,} \\ \nu^2 & \text{for NLKG.} \end{cases}$$

We note that by virtue of (1.3), the nonlinearity $\vec{f}$ has the analogous equivariant property

$$\vec{f}(y) = f(|y|)\hat{y} \qquad \text{for} \quad y \neq 0, \tag{1.6}$$

where the scalar-valued function f is defined by $f(s) \equiv g(s) + \omega s$. For convenience, we will also denote by f the odd extension of f to all of $\mathbb{R}$.

In this paper, we prove the existence of twice-differentiable solutions $v : \mathbb{R}^N \to \mathbb{R}^M$ with $N \geq 2$ and $M \geq 2$ such that $\nu(x) \to 0$ as $|x| \to \infty$. We remark that if v is such a solution of (1.5), then $u(x,t) \equiv e^{\nu t K}\, v(x)$ is a localized standing-wave solution of the corresponding nonlinear evolution equation (NLS) or (NLKG). Under a (Galilean or Lorentz) velocity boost, such a standing wave becomes a spatially-localized traveling wave solution of the wave equation, that is, a multidimensional solitary wave.

The set of spherically-symmetric ("radial") scalar-valued ($M = 1$) solutions to (1.5) has been extensively studied (see [1], [2], [3], [5], [6], [8], [9], [10], [12], [13], [14], [15]). Constructive techniques to establish the existence of localized solutions to (1.5) with prescribed nodal structure include [13], which treats real-valued radial solutions in $\mathbb{R}^N$ for nonlinearities that are superlinear at large amplitudes.

In [7] we used techniques analogous to those in [13] to investigate the existence and nodal structure of complex-valued ($M = 2$) nonradial solutions in $\mathbb{R}^2$ for nonlinearities that are superlinear at large amplitudes. The corresponding standing-wave solutions to (NLS) and (NLKG) carry nonzero (classical) angular momentum in the center-of-momentum frame and are thus "spinning" solitary waves.

The type of nonlinearity treated in [7] and [13], however, admits finite-time blowup of (some) solutions to the corresponding semilinear wave equation. In [8], we studied a different type of nonlinearity, compatible with global regularity of solutions, in which the energy density for the associated wave equation is positive definite. For such "hilltop" nonlinearities, [8] establishes the existence, in arbitrary spatial dimension, of spherical (real-valued, spinless) solitary-wave solutions to (NLS) and (NLKG) with a prescribed nodal structure. Here we establish the existence, in arbitrary spatial dimension, of nonspherical (spinning) solitary-wave solutions with prescribed nodal structure, for "hilltop" nonlinearities compatible with global regularity of solutions.

For complex-valued nonradial solutions in spatial dimensions $N \geq 3$, no ansatz has been found that reduces (1.5) to an ordinary differential equation. An ansatz for complex-valued nonradial solutions that transforms (1.5) to a partial differential equation with fewer independent variables is made in [11], where the existence of localized solutions is established through variational methods. In contrast, the vector-valued solutions we construct here (with range dimension $M \geq 3$ in the case $N \geq 3$) result from an ansatz that reduces (1.5) to an ordinary differential equation.

In particular, we look for solutions to the elliptic equation (1.5) of the separated form

$$v(x) = w(r)\, \hat{\psi}(\hat{x}), \tag{1.7}$$

where $r \equiv |x|$ and $\hat{x} \equiv \frac{x}{|x|}$. Here $w : [0,\infty) \to \mathbb{R}$, and $\hat{\psi} : S^{N-1} \to S^{M-1}$ is a function on the sphere S^{N-1} in $\mathbb{R}^N$, taking unit-vector values in $\mathbb{R}^M$. For $v(x)$ of this form, we have

$$\Delta v = (\Delta_r w)\hat{\psi} + \tfrac{1}{r^2} w\,(\Delta_S \hat{\psi}), \tag{1.8}$$

where $\Delta_r \equiv \frac{\partial^2}{\partial r^2} + \frac{N-1}{r}\frac{\partial}{\partial r}$ is the radial Laplacian, and where Δ_S is the Laplacian on the sphere $S^{N-1} \subset \mathbb{R}^N$, applied componentwise to $\hat{\psi}$.

We show in section 2 that there are functions $\hat{\psi} : S^{N-1} \to S^{M-1}$ whose coordinate functions are eigenfunctions of the spherical Laplacian with eigenvalue $\mu_\ell \equiv -\ell(\ell + N - 2)$, provided that the dimension M of the range space is at least as large as the multiplicity of μ_ℓ. Given that

$$\Delta_S \hat{\psi} = -\ell(\ell + N - 2)\hat{\psi}, \tag{1.9}$$

it follows from (1.5)–(1.9) that the radial profile w must then satisfy the ordinary differential equation

$$w'' + \frac{N-1}{r} w' - \frac{\ell(\ell + N - 2)}{r^2} w + f(w) = 0\,. \tag{1.10}$$

We use ordinary-differential-equation arguments to prove that, for each prescribed number of nodes, there is a solution $w : [0,\infty) \to \mathbb{R}$ to (1.10) that is exponentially decreasing far from the origin. We establish this result for $N \geq 2$ and $\ell \geq 1$, under hypotheses on the nonlinearity f that are compatible with global existence of finite-energy solutions to the associated nonlinear wave equations, as in [8]. For our nonlinearities, the corresponding standing-wave solutions for (NLS) have a structure consisting of a core region resembling a vortex (a nonlocalized solution with asymptotically constant amplitude), encapsulated by a membrane region that exponentially localizes the solution.

Let G be the primitive of g. The individual terms in the conserved energy densities $e_1[u] \equiv \frac{1}{2}|\nabla u|^2 - G(|u|)$ for (NLS) and $e_2[u] \equiv \frac{1}{2}|u_t|^2 + \frac{1}{2}|\nabla u|^2 - G(|u|)$ for (NLKG) are all nonnegative provided that G is nonpositive. In that case, the time-independence of the spatial integrals of $e_1[u]$ and $e_2[u]$ ensures global existence of finite-energy solutions. Because $f(s) = g(s) + \omega s$, the primitive F of f is related to G by $F(s) = G(s) + \frac{1}{2}\omega s^2$. Accordingly, we treat nonlinearities f in (1.10) whose primitives are quadratically bounded above, corresponding to nonpositive G.

Our hypotheses on f are essentially those for the "hilltop case" in [8]. We suppose that f is an odd locally Lipschitz-continuous function with $-\infty < -\sigma^2 \equiv \lim_{s\to 0} \frac{f(s)}{s} \leq 0$, and in case $\sigma = 0$ we require that $f(s) < 0$ for small positive s. We assume that there exist β and δ with $0 < \beta < \delta$ such that $f(\beta) = f(\delta) = 0$ and $f > 0$ on (β, δ). For simplicity, we assume that $f(s)$ is strictly decreasing for s near δ. In addition, we assume that the primitive $F(s) \equiv \int_0^s f(t)\,dt$ has a positive zero γ with $\beta < \gamma < \delta$ and that $F < 0$ on $(0, \gamma)$. We denote the smallest positive zero of f by α. Thus $0 < \alpha \leq \beta < \gamma < \delta$. See Figure 1. We make no assumptions on $f(s)$ for arguments with $|s| > \delta$.

To formulate a well-posed initial value problem for (1.10) with initial conditions given at $r = 0$, we make the change of variable $w(r) = r^\ell z(r)$ to obtain the equation

$$z'' + \tfrac{2\ell+N-1}{r} z' + \tfrac{1}{r^\ell} f(r^\ell z) = 0 \tag{1.11}$$

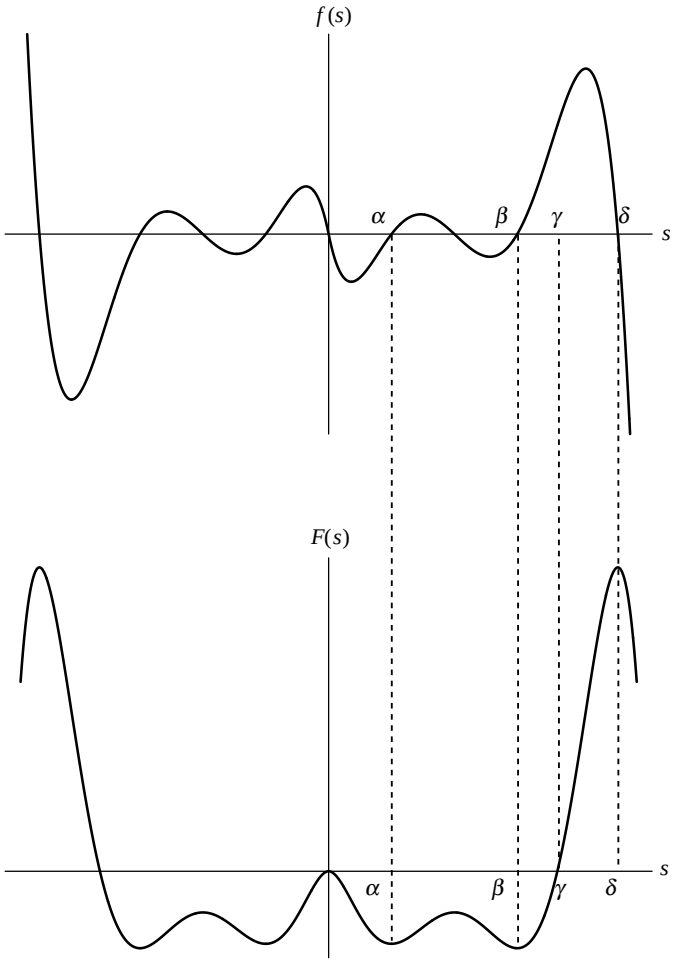


FIG. 1.

for z, which, by virtue of the condition on f at zero, has well-posed initial value problems obtained by specifying

$$z(0) = d \text{ and } z'(0) = 0. \tag{1.12}$$

Since f is odd it suffices to consider nonnegative d, and henceforth we assume $d \geq 0$. If z is a C^2 solution of such an initial value problem, it follows that $w(r) = r^\ell z(r)$ is a C^2 solution of the differential equation

$$w'' + \frac{N-1}{r} w' - \frac{\ell(\ell+N-2)}{r^2} w + f(w) = 0 \tag{1.13}$$

subject to

$$\lim_{r \to 0^+} \frac{w(r)}{r^\ell} = d \quad \text{and} \quad \lim_{r \to 0^+} \left(\frac{w(r)}{r^\ell} \right)' = 0 \tag{1.14}$$

and that the corresponding function $v(x) = w(|x|)\hat{\psi}(\hat{x})$ is C^2 on $\mathbb{R}^N$. We remark that the term $\frac{\ell(\ell+N-2)}{r^2} w$ in (1.13) makes the behavior of solutions significantly different

from that in the "radial" case $\ell = 0$. See [7] and [8] for commentary on the novel features of these initial value problems.

We remark that the equivalence of (1.10) and (1.11) shows that, for the case of superlinearly-growing f (not under consideration here), the analysis in [7] for the $N = 2$ case applies to establish the existence of localized standing-wave solutions to (NLS) and (NLKG) for such f and general N.

We prove the following main theorem.

MAIN THEOREM. *Let the nonlinearity f have the ("hilltop") properties specified. Let integers $N \geq 2$ and $\ell \geq 1$ be given. Then, for each nonnegative integer n, there is a positive number d and a C^2 solution w to* (1.13) *and* (1.14) *such that* $\lim_{r\to\infty} w(r) = 0$ *and w has exactly n positive zeros.*

Figure 2 shows the radial profiles $w(r)$ with $n = 0$ and $n = 1$ positive zeros in the case $N = 2$ and $\ell = 2$ for $f(s) \equiv -(4.3)s + 5s^3 - s^5$.

We remark that if $\sigma \neq 0$, then the solutions whose existence is established by the main theorem decay exponentially as $r \to \infty$. This can be established using the proof technique in section 6 of [7].

In section 2, we show that there are unit-vector-valued eigenfunctions of the Laplacian on the sphere. In section 3, we return to consideration of the ordinary differential equation and establish some general properties of solutions of (1.13)–(1.14). In section 4, we show that there is a value of d for which the solution to (1.13)–(1.14) is monotonically increasing to δ as $r \to \infty$. Section 5 contains the proof of the main theorem.

In the following, we write $r \to 0$ and $d \to 0$ to mean $r \to 0^+$ and $d \to 0^+$, respectively. Throughout, $N \geq 2$, $\ell \geq 1$, and $w(r) = r^\ell z(r)$.

We will make use of the identity

$$r^2(\tfrac{1}{2}w'^2 + F(w))|_{r_1}^{r_2} + (N-2)\int_{r_1}^{r_2} sw'^2\,ds = \tfrac{\ell(\ell+N-2)}{2}w^2|_{r_1}^{r_2} + 2\int_{r_1}^{r_2} sF(w(s))\,ds \tag{1.15}$$

which results from multiplying (1.13) by r^2w' to obtain $(\frac{1}{2}r^2w'^2)' + (N-2)rw'^2 - (\frac{\ell(\ell+N-2)}{2}w^2)' + r^2(F(w))' = 0$, and then integrating on (r_1, r_2).

2. Vector-valued eigenfunctions of the Laplacian on the sphere. In this section we show that unit-vector-valued eigenfunctions of the Laplacian on the sphere exist for suitably large dimension M of the range space. We first recall [4, section 2H] that the Laplacian Δ_S on the unit sphere S^{N-1} in $\mathbb{R}^N$ has scalar-valued eigenfunctions with eigenvalues $\mu_\ell \equiv -\ell(\ell+N-2)$, $\ell = 0,\ 1,\ 2,\ \ldots$; the subspace H_ℓ of $L^2(S^{N-1})$ consisting of eigenfunctions with eigenvalue μ_ℓ has dimension

$$D_\ell = \frac{(2\ell+N-2)(\ell+N-3)!}{(N-2)!\,\ell!}. \tag{2.1}$$

(In the case $N = 2$, we have $D_0 = 1$ and $D_\ell = 2$ for all $\ell \geq 1$.) For given N and ℓ, let $\{Y_{\ell m} \mid m = 1,\ 2,\ \ldots,\ D_\ell\}$ be an orthonormal basis for H_ℓ consisting of real-valued functions. Then the identity

$$\sum_{m=1}^{D_\ell} (Y_{\ell m}(\hat{x}))^2 = \frac{D_\ell}{\sigma_N} \tag{2.2}$$

holds for all $\hat{x} \in S^{N-1}$, where $\sigma_N \equiv \frac{2\pi^{N/2}}{\Gamma(N/2)}$ is the area of S^{N-1} in $\mathbb{R}^N$.

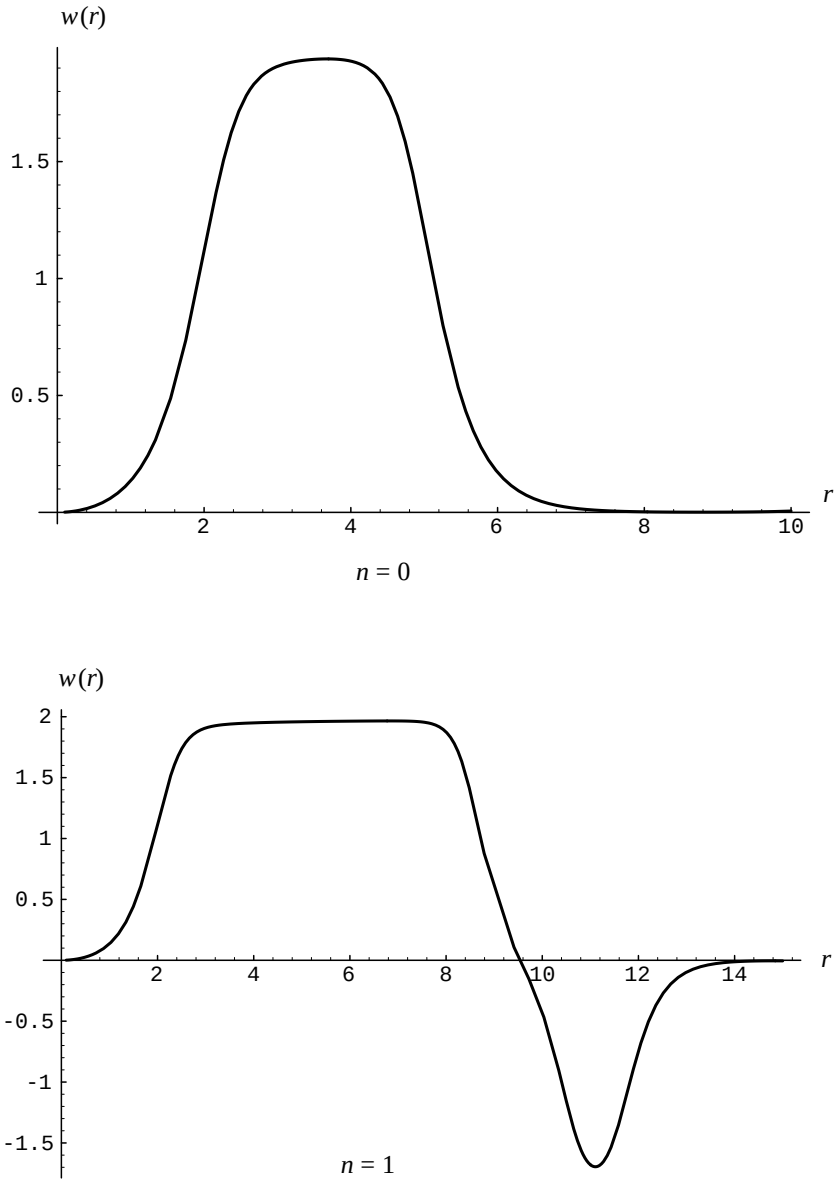


FIG. 2. *Radial profiles $w(r)$ in the case $N = 2$ and $l = 2$ for $f(s) = -(4.3)s + 5s^3 - s^5$.*

We now construct unit-vector-valued eigenfunctions. Given the spatial dimension N and a nonnegative integer ℓ, a vector-valued eigenfunction of Δ_S with eigenvalue μ_ℓ has the general form

$$\psi(\hat{x}) \equiv \sum_{m=1}^{D_\ell} a_m Y_{\ell m}(\hat{x}), \tag{2.3}$$

where each a_m is a (constant) vector in $\mathbb{R}^M$. The requirement that ψ be a unit vector

for all $\hat{x} \in S^{N-1}$ then becomes

$$1 = |\psi(\hat{x})|^2 = \sum_{m=1}^{D_\ell} \sum_{n=1}^{D_\ell} (a_m \cdot a_n)\, Y_{\ell m}(\hat{x}) Y_{\ell n}(\hat{x}), \tag{2.4}$$

where $a_m \cdot a_n$ is the ordinary Euclidean inner product of the vectors a_m and a_n. In view of the identity (2.2), we may satisfy this requirement by choosing the vectors $\{a_m \mid m = 1,\ 2,\ \ldots,\ D_\ell\}$ to be pairwise orthogonal and normalized by $|a_m| = \sqrt{\sigma_N / D_\ell}$.

We note that in order to accommodate the D_ℓ linearly independent vectors in this construction, the dimension M of the range space must be at least D_ℓ. In the case of $N = 2$ spatial dimensions, $D_\ell = 2$ for all $\ell \geq 1$, and the maps $\psi_\ell : S^1 \to S^1$ given by $\psi_\ell(\theta) \equiv (\cos\ell\theta, \sin\ell\theta)$ are eigenfunctions of Δ_S with eigenvalue $\mu_\ell = -\ell^2$. This is the situation treated in [7] (for growing nonlinearity f).

In higher spatial dimensions, D_ℓ is an increasing function of ℓ, so that the range of admissible values of ℓ in this construction depends on M. In particular, since our construction requires D_ℓ linearly independent vectors in $\mathbb{R}^M$, for fixed domain dimension N and range dimension M, there is a maximum value $\ell_{\max} = \max\{\ell : D_\ell \leq M\}$ of ℓ for which our construction can produce unit-vector-valued eigenfunctions of Δ_S with eigenvalue μ_ℓ. (This upper limit on the "spin" ℓ is of potential interest for the modeling of elementary particles with equations of these types.)

We note, however, that $D_1 = N$ for all N, so that a vector field ($M = N$) accommodates eigenvalue μ_1. In particular, $\psi_1 : S^{N-1} \to S^{N-1}$, defined by $\psi_1(\hat{x}) \equiv \hat{x}$, is an eigenfunction of Δ_S with eigenvalue $\mu_1 = -(N-1)$ for general N.

In all cases, given spatial dimension N and index ℓ, we may choose suitably-normed pairwise orthogonal vectors a_m so that (2.3) furnishes unit-vector-valued eigenfunctions of the Laplacian on S^{N-1} with eigenvalue μ_ℓ, provided that the dimension M of the range space is sufficiently large to accommodate D_ℓ linearly independent vectors.

3. Properties of solutions to the initial value problem. As in [7], it is a straightforward task to show that for any fixed $d \geq 0$, the initial value problem (1.13)–(1.14) has a unique C^2 solution on some small interval $(0, \epsilon)$ with $\epsilon > 0$. The solution $w(r) = w(r, d)$ of (1.13)–(1.14) may be continued in $r > 0$ as long as $|w(r)| \leq \delta$ and $|w'(r)|$ remains finite. Let $(0, R(d))$ be the resulting maximal open interval of existence for $w(r, d)$. If $R(d) = \infty$, then $|w(r, d)| \leq \delta$ for $r \in (0, \infty)$. If $0 < R(d) < \infty$, then, because f is bounded on $[-\delta, \delta]$, both limits $w(R(d), d) \equiv \lim_{r \to R(d)^-} w(r, d)$ and $w'(R(d), d) \equiv \lim_{r \to R(d)^-} w'(r, d)$ exist, which we can verify as follows.

For $r \in (0, R(d))$, we have $r^\ell |z| = |w| \leq \delta$. Multiplying (1.11) by $r^{2\ell+N-1}$ and integrating give

$$-z'(r, d) = \frac{1}{r^{2\ell+N-1}} \int_0^r s^{\ell+N-1} f(s^\ell z(s, d))\, ds. \tag{3.1}$$

Since f is bounded on $[-\delta, \delta]$, we see that $|z'(r, d)| \leq (\text{constant}) r^{1-\ell}$ for $r \in (0, R(d))$. Thus, $\lim_{r \to R(d)^-} z(r, d)$ exists. Furthermore, because f is bounded and $|z'|$ is bounded for r near $R(d)$, (1.11) shows that $|z''(r, d)|$ is also bounded for r near $R(d)$, and thus $\lim_{r \to R(d)^-} z'(r, d)$ exists as well. Because $w(r, d) = r^\ell z(r, d)$, it follows that the limits of w and w' as $r \to R(d)^-$ both exist, as asserted.

We now establish some qualitative properties of solutions $w(r)$ to (1.13)–(1.14).

LEMMA 3.1. *Suppose p is a local maximum for $|w|$ with $0 < p < R(d)$ and $0 < |w(p)| < \delta$. Then $F(w(r)) \le F(w(p))$ for $r > p$. In particular, if $|w(p)| \ge \gamma$, then $|w(r)| \le |w(p)|$ for $r > p$.*

Proof. Consider first the case $0 < w(p) < \delta$. Since p is a local maximum for w, we have that $w'(p) = 0$ and $w''(p) \le 0$. Therefore, from (1.13) we have $f(w(p)) \ge \frac{\ell(\ell+N-2)}{p^2} w(p) > 0$. Thus, F is increasing near $w(p)$. Since $w(r) \le w(p)$ for r near p, we then have that $F(w(r)) \le F(w(p))$ for r near p.

Next, suppose by way of contradiction that the lemma is false. Then there exists a finite number s with $p < s$ such that $[p, s]$ is the r-interval of maximal length on which $F(w(r)) \le F(w(p))$. It follows that $F(w(s)) = F(w(p))$.

We now evaluate the identity (1.15) with $r_1 = p$, $r_2 = s$. This gives

$$F(w(p))[s^2 - p^2] \le \tfrac{\ell(\ell+N-2)}{2}[w^2(s) - w^2(p)] + 2\int_p^s tF(w)\,dt. \tag{3.2}$$

Since $F(w(r)) \le F(w(p))$ on $[p, s]$, we have

$$2\int_p^s tF(w)\,dt \le F(w(p))[s^2 - p^2].$$

Thus (3.2) implies

$$w^2(p) \le w^2(s).$$

Therefore, there exists q with $p < q \le s$ such that $w(p) = |w(q)|$. Again, using identity (1.15), this time with $r_2 = q$ and $r_1 = p$, gives (since F is even and thus $F(w(q)) = F(w(p))$)

$$\frac{1}{2}q^2 w'(q)^2 + F(w(p))[q^2 - p^2] \le 2\int_p^q rF(w)\,dr. \tag{3.3}$$

We may assume that $F(w(r)) \not\equiv F(w(p))$ on $[p, q]$. (Otherwise, $w(r)$ would be constant, which is impossible). Thus (3.3) implies

$$\frac{1}{2}q^2 w'(q)^2 + F(w(p))[q^2 - p^2] \le 2\int_p^q rF(w)\,dr < F(w(p))[q^2 - p^2],$$

or equivalently, $q^2 w'(q)^2 < 0$, which is absurd. This contradiction completes the proof of Lemma 3.1 in the case where $w(p)$ is positive.

To treat the situation in which $w(p)$ is negative, we note that because f is odd, $\tilde{w}(r) \equiv -w(r)$ is also a solution to (1.13), and $\tilde{w}(p) > 0$ is a local maximum of $\tilde{w}$. Thus our arguments apply without change to $\tilde{w}$, and complete the proof of Lemma 3.1 in this case as well.

LEMMA 3.2. *Suppose that w is not identically zero and there exists $t \in (0, R(d))$ such that $w(t) = 0$. Then, there exists q with $0 < q < t$ such that $|w(q)| > \gamma$.*

Proof. The quantity $w'(t)^2$ is strictly positive, since otherwise $w \equiv 0$ by uniqueness of solutions of initial value problems. Using identity (1.15) with $r_2 = t$ and $r_1 = 0$, we thus obtain

$$0 < \frac{1}{2}t^2 w'(t)^2 \le 2\int_0^t sF(w(s))\,ds.$$

Thus the integrand is positive somewhere, that is, $F(w(q)) > 0$ for some $q \in (0,t)$. Hence $|w(q)| > \gamma$, as claimed.

LEMMA 3.3. *Suppose $w'(r) \geq 0$ on $(a,b) \subset (0,R(d))$, $w(a) \geq 0$, and $w \not\equiv 0$. Then in fact $w'(r) > 0$ on (a,b).*

Proof. Suppose there exists $p \in (a,b)$ with $w'(p) = 0$. Then $w(p) \geq w(a) \geq 0$ since $w' \geq 0$ on (a,b). In addition, $w(p) > 0$ because otherwise $w \equiv 0$ by the uniqueness of solutions of initial value problems. Now, since $w' \geq 0$ on (a,b), $w'(r)$ has a minimum at p, and therefore $w''(p) = 0$. We next observe that the composition $f(w(r))$ is differentiable at $r = p$ even though f is only locally Lipschitz, and $\frac{d}{dr} f(w(r))|_{r=p} = 0$. To see this, observe that $0 \leq \lim_{r\to p} |\frac{f(w(r))-f(w(p))}{r-p}| \leq K \lim_{r\to p} |\frac{w(r)-w(p)}{r-p}| = K|w'(p)| = 0$, where K is the Lipschitz constant for f near $w(p)$. Thus we may differentiate (1.13) at $r = p$, to obtain $w'''(p) = -\frac{2\ell(\ell+N-2)}{p^3} w(p) < 0$. But $w'''(p)$ cannot be negative if $w'(r)$ has a minimum at $r = p$. Thus there cannot be any point $p \in (a,b)$ with $w'(p) = 0$, which establishes the assertion of Lemma 3.3.

LEMMA 3.4. *If $w(q) = \delta$ for some finite $q \in (0,R(d)]$, then $w'(q) > 0$.*

Proof. Recall that $|w(r)| \leq \delta$ for $r \in (0,R(d)]$. Thus $w'(q) \geq 0$. Now suppose $w'(q) = 0$. Then $w''(q) \leq 0$. On the other hand, from (1.13) we have $w''(q) = \frac{\ell(\ell+N-2)}{q^2}\delta > 0$, a contradiction. Thus, $w'(q) > 0$ and the lemma is proved.

LEMMA 3.5. *$w(r) > -\delta$ for all $r \in (0,R(d))$. If $R(d) < \infty$, then also $w(R(d)) > -\delta$.*

Proof. Recall that $d \geq 0$, so that (from (1.14)) w is positive for small positive r. If there exists $r_0 \in (0,R(d)]$ with $w(r_0) = -\delta$, then w must have a local maximum, p, with $0 < p < r_0$ and $0 < w(p) \leq \delta$. By Lemma 3.4, $0 < w(p) < \delta$. Now by Lemma 3.1, $F(w(r)) \leq F(w(p)) < F(\delta)$ for $r > p$. Evaluating at r_0 gives $F(\delta) < F(\delta)$, a contradiction. The lemma is proved.

We now define

$$\tilde{E}(r) = \frac{\frac{1}{2}w'^2(r) + F(w(r)) - F_0}{r^{2\ell-2}} + \frac{\ell(\ell+N-2)}{2}\frac{w^2(r)}{r^{2\ell}}, \tag{3.4}$$

where

$$F_0 \equiv \min_{y\in[-\delta,\delta]} F(y) < 0. \tag{3.5}$$

We note that $\tilde{E}(r) \geq 0$, and from (1.14) we see that

$$\tilde{E}(0) = +\infty \quad \text{if} \quad \ell > 1$$

and

$$\tilde{E}(0) = \frac{N}{2}d^2 - F_0 > 0 \quad \text{if} \quad \ell = 1.$$

A computation shows that

$$\tilde{E}' = -\frac{(\ell+N-2)}{r^{2\ell-1}}\left(w' - \frac{\ell w}{r}\right)^2 - \frac{2(\ell-1)[F(w(r)) - F_0]}{r^{2\ell-1}} \leq 0. \tag{3.6}$$

For each solution $w(r)$, $\tilde{E}(r)$ is thus a nonnegative, nonincreasing quantity. Since the term $F(w(r)) - F_0$ is nonnegative, we have that

$$\frac{w'^2}{r^{2\ell-2}} + \frac{w^2}{r^{2\ell}} \leq \tilde{E}(r_0) < \infty \tag{3.7}$$

for $0 < r_0 < r < R(d)$. From this estimate and Lemmas 3.4 and 3.5, it follows that the only way in which a solution $w(r)$ can fail to extend to all positive r is if $w(r)$ attains the value δ. Furthermore, if $R(d) < \infty$ then $|w| < \delta$ on $[0, R(d))$ and $w(R(d)) = \delta$. Otherwise, if $R(d) = \infty$, then $|w| < \delta$ on $[0, \infty)$.

LEMMA 3.6. *Let $w(r, d)$ be the solution of the initial value problem* (1.13)–(1.14). *If $R(d) < \infty$, then $w'(r, d) > 0$ on $(0, R(d)]$.*

Proof. If $d = 0$, then w is identically zero, so then $R(d) = \infty$. Thus under the hypotheses here, $d > 0$. From (1.12) and the relation $w(r) = r^\ell z(r)$, it follows that $w'(r) > 0$ for small $r > 0$. Now suppose that w has a local maximum, p, with $0 < p < R(d)$ and $0 < w(p) < \delta$. By Lemma 3.1, $F(w(r)) \le F(w(p)) < F(\delta)$ for $r > p$. Evaluating at $R(d)$ gives $F(\delta) < F(\delta)$, a contradiction. Hence, $w'(r) \ge 0$ on $(0, R(d))$. Then, by Lemma 3.3, $w'(r) > 0$ on $(0, R(d))$, and by Lemma 3.4, $w'(R(d)) > 0$. This completes the proof of the lemma.

4. Existence of a positive, monotone solution with limiting value δ. Recall that $\ell \ge 1$ and $N \ge 2$. The proofs of Lemmas 4.1 and 4.2 below are implicit in [7, section 3].

LEMMA 4.1. *For $d > 0$ chosen small enough,* (1.13)–(1.14) *has a solution, $w(r, d)$, which satisfies $0 < w(r, d) < \gamma$ for all $r > 0$.*

LEMMA 4.2. *For any $d > 0$, there exists a smallest value of r, a_d, such that $w(a_d, d) = \alpha$, and $w' > 0$ for all $r \in (0, a_d)$.*

LEMMA 4.3. *For $r \in (0, a_d)$ we have $w(r) \ge dr^\ell$. (Consequently, taking limits as $r \to a_d^-$ gives $\alpha \ge da_d^\ell$. Therefore, $a_d \to 0$ as $d \to \infty$).*

Proof. On $[0, a_d]$ we have $f(w) \le 0$. Hence, from (1.11) we obtain

$$(r^{2\ell+N-1} z')' \ge 0.$$

Integrating on $(0, r)$ gives $z' \ge 0$, that is, $(w(r)/r^\ell)' \ge 0$. Thus, using (1.14) we have

$$w(r) \ge dr^\ell$$

on $(0, a_d)$. This completes the proof of Lemma 4.3.

LEMMA 4.4. *For $d > 0$ large enough, we have $R(d) < \infty$.*

Proof. Assume by way of contradiction that $R(d) = \infty$ for an unbounded sequence of positive values of d. For such values of d with $R(d) = \infty$, we know by Lemma 3.4 that $|w| < \delta$ on $(0, \infty)$. Then (1.13) gives

$$w'' + \frac{N-1}{r} w' - \frac{\ell(\ell+N-2)}{r^2} w + M \ge 0,$$

where $M \equiv \max_{y \in [-\delta, \delta]} f(y)$. Letting $w = r^\ell z$ yields

$$(r^{2\ell+N-1} z')' \ge -M r^{\ell+N-1}.$$

Hence, integrating on $(0, r)$ and using (1.12) gives

$$z' \ge -\frac{M}{\ell+N} r^{1-\ell}. \tag{4.1}$$

We now must consider three different cases.

Case I ($\ell = 1$). Integrating (4.1) on $(0, r)$ and using $z(0) = d$ give

$$z \ge d - \frac{M}{1+N} r$$

and thus

$$w \geq r\left(d - \frac{M}{1+N}r\right). \tag{4.2}$$

Evaluating at $r = 1$ gives

$$w(1) \geq d - \frac{M}{1+N}. \tag{4.3}$$

Clearly, for large enough d, we have $w(1) > \delta$, which is a contradiction to our assumption that $R(d) = \infty$ (for this implied that $|w| < \delta$). Thus, we must have that $R(d) < \infty$ for large enough d.

This completes the lemma for Case I.

Case II ($\ell = 2$). We begin by integrating (4.1) on (a_d, r) and we obtain

$$z(r) \geq z(a_d) - \frac{M}{2+N}[\log r - \log a_d] \quad \text{for} \quad r > a_d.$$

Since $w = r^2 z$, we have

$$\frac{w}{r^2} \geq \frac{\alpha}{a_d^2} + \frac{M}{2+N}\log a_d - \frac{M}{2+N}\log r \quad \text{for} \quad r > a_d.$$

By Lemma 4.3, $a_d \to 0$ as $d \to \infty$. Thus, $a_d < 1$ for large enough d. Evaluating the above inequality at $r = 1$ we obtain

$$w(1) \geq \frac{\alpha}{a_d^2} + \frac{M}{2+N}\log a_d = \frac{1}{a_d^2}\left(\alpha + \frac{M}{2+N}a_d^2 \log a_d\right).$$

Since $a_d \to 0$ as $d \to \infty$, it follows that $\lim_{d\to\infty} a_d^2 \log a_d = 0$. So, for large enough d, we have $w(1) > \delta$ which again is a contradiction to the fact that $R(d) = \infty$. Thus, $R(d) < \infty$ for large enough d. This completes Case II.

Case III ($\ell > 2$). We begin by integrating (4.1) on (a_d, r) and obtain

$$\frac{w}{r^\ell} = z(r) \geq z(a_d) + \frac{M}{(\ell+N)(\ell-2)}\left[\frac{1}{r^{\ell-2}} - \frac{1}{a_d^{\ell-2}}\right] = \frac{\alpha}{a_d^\ell} + \frac{M}{(\ell+N)(\ell-2)}\left[\frac{1}{r^{\ell-2}} - \frac{1}{a_d^{\ell-2}}\right]$$

$$= \frac{1}{a_d^{\ell-2}}\left[\frac{\alpha}{a_d^2} - \frac{M}{(\ell+N)(\ell-2)}\right] + \frac{M}{(\ell+N)(\ell-2)}\frac{1}{r^{\ell-2}}. \tag{4.4}$$

Because $a_d \to 0$ as $d \to \infty$, the term in brackets in (4.4) is positive if d is chosen large enough. Thus for sufficiently large d,

$$w \geq \frac{M}{(\ell+N)(\ell-2)}r^2 \quad \text{for } r > a_d. \tag{4.5}$$

Thus $w > \delta$ for large enough r, which contradicts the assumption that $R(d) = \infty$. Therefore, $R(d) < \infty$. This completes Case III and thus completes the proof of Lemma 4.4.

We now let D be the set of d-values for which the solution reaches the hilltop at a finite time. That is, we define

$$D = \{d > 0 | R(d) < \infty\}.$$

By Lemma 4.4, if d is chosen large enough, then $R(d) < \infty$. Thus the set D is nonempty. Also, by Lemma 4.1, we have $0 < w < \gamma < \delta$ for all $r > 0$ if d is chosen small enough. Thus, D is bounded away from zero.

We now let

$$d^* = \inf D.$$

Note that $d^* > 0$.

LEMMA 4.5. *For $d \in (0, d^*)$, the solution $z(r,d)$ of (1.11)–(1.12) satisfies*

$$|z(r,d)| \le C_1, \qquad |z'(r,d)| \le C_2, \quad \text{and} \quad |z''(r,d)| \le C_3$$

for all $r > 0$, where C_1, C_2, and C_3 are independent of d.

Proof. Since $d \in (0, d^*)$, in particular $d \notin D$. Thus $R(d) = \infty$ which implies (by Lemma 3.4) that

$$|w(r,d)| < \delta \ \text{ on } \ [0,\infty). \tag{4.6}$$

We now estimate $z(r,d)$. Integrating (3.1) gives

$$z(r,d) = d - \int_0^r \frac{\int_0^t s^{\ell+N-1} f(s^\ell z(s,d))\,ds}{t^{2\ell+N-1}}\,dt. \tag{4.7}$$

Next, note that $\frac{f(u)}{u}$ is bounded (since f is bounded and $\lim_{u\to 0}\frac{f(u)}{u} = -\sigma^2$) so that there is a $B > 0$ such that $|f(u)| \le B|u|$. Now let $s_0 \in [0, \sqrt{(2\ell+N)/B}]$ be such that

$$|z(s_0,d)| = \max_{\left[0,\sqrt{\frac{2\ell+N}{B}}\right]} |z(r,d)|.$$

Estimating (4.7) at s_0 gives

$$|z(s_0,d)| \le d + B|z(s_0,d)| \int_0^{s_0} \frac{\int_0^t s^{2\ell+N-1}\,ds}{t^{2\ell+N-1}}\,dt = d + \frac{B|z(s_0,d)|s_0^2}{2(2\ell+N)} \le d + \frac{1}{2}|z(s_0,d)|.$$

Consequently,

$$\max_{\left[0,\sqrt{\frac{2\ell+N}{B}}\right]} |z(r,d)| \le 2d \le 2d^*. \tag{4.8}$$

Further, by (4.6) we know that

$$\max_{\left[\sqrt{\frac{2\ell+N}{B}},\infty\right)} |z(r,d)| = \max_{\left[\sqrt{\frac{2\ell+N}{B}},\infty\right)} \frac{|w(r,d)|}{r^\ell} < \delta\left(\frac{B}{2\ell+N}\right)^{\ell/2}. \tag{4.9}$$

Combining (4.8) and (4.9), we then have for $0 < d < d^*$,

$$|z(r,d)| \le 2d^* + \delta\left(\frac{B}{2\ell+N}\right)^{\ell/2} = C_1(d^*, \delta, \ell, N, B).$$

Next, we estimate $z'(r,d)$. Equation (3.1) gives

$$|z'(r,d)| = \left|\frac{1}{r^{2\ell+N-1}}\int_0^r s^{\ell+N-1} f(s^\ell z(s,d))\,ds\right| \le \frac{BC_1}{2\ell+N} r. \tag{4.10}$$

On the other hand, we may obtain a different estimate for $|z'|$ by using the fact that $|f(y)| \le B\delta$ for $y \in [-\delta, \delta]$. Equation (3.1) gives

$$|z'(r,d)| \le \frac{B\delta}{r^{2\ell+N-1}} \int_0^r s^{\ell+N-1}\, ds = \left(\frac{B\delta}{\ell+N}\right) \frac{1}{r^{\ell-1}}. \tag{4.11}$$

Combining the two estimates (4.10) and (4.11) by locating the intersection of the graphs of the upper bounds as functions of r, we obtain

$$|z'(r,d)| \le B\left(\frac{\delta}{\ell+N}\right)^{1/\ell} \left(\frac{C_1}{2\ell+N}\right)^{(\ell-1)/\ell} = C_2(d^*, \delta, \ell, N, B).$$

Last, using (1.11) and the previous estimates, we obtain

$$|z''(r,d)| \le C_3(d^*, \delta, \ell, N, B).$$

This completes the proof of Lemma 4.5.

LEMMA 4.6. *$R(d^*) = \infty$.*

Proof. For $d \in (0, d^*)$, Lemma 4.5 establishes that all of $|z(r,d)|$, $|z'(r,d)|$, and $|z''(r,d)|$ are bounded uniformly in d for all $r \ge 0$. Thus, by the Arzelá–Ascoli theorem, there exists a sequence $\{d_j\}$ tending to d^* from below and a function $z^*(r)$ such that $z(r,d_j) \to z^*(r)$ and $z'(r,d_j) \to z^{*\prime}(r)$ as $j \to \infty$ (where the convergence is uniform on compact sets). In addition, taking limits in (1.11) as $d \to d^{*-}$ shows that z^* solves (1.11) on $[0,\infty)$.

Now suppose that the assertion of the lemma is false, that is, assume that $R(d^*) < \infty$. By uniqueness of solutions to initial value problems, we know that $z^*(r)$ is equal to $z(r,d^*)$ on $[0, R(d^*)]$. Thus, $w^*(r) \equiv r^\ell z^*(r)$ satisfies (1.13)–(1.14) on $[0,\infty)$ and $w^*(r) = w(r,d^*)$ on $[0,R(d^*)]$. We also have from (4.6) that $|w(r,d)| < \delta$ for $0 \le d < d^*$. Taking the limit as $d \to d^{*-}$, we see that $|w^*(r)| \le \delta$ for all $r > 0$. Since $w^*(R(d^*)) = w(R(d^*), d^*) = \delta$, we see that $w^*(r)$ has a local maximum at $R(d^*)$. This contradicts the fact that $w^{*\prime}(R(d^*)) = w'(R(d^*), d^*) > 0$, established by Lemma 3.4. Thus we must have $R(d^*) = \infty$. This completes the proof of Lemma 4.6.

LEMMA 4.7. *$w(r,d^*) > 0$ and $w'(r,d^*) > 0$ for $r > 0$, and $\lim_{r\to\infty} w(r,d^*) = \delta$.*

Proof. We adopt the shorthand $w^*(r) \equiv w(r,d^*)$. Since $R(d^*) = \infty$, we know that $|w^*(r)| < \delta$ for all $r \ge 0$. Also, since $d^* > 0$, from (1.14) we know that $w^*(r)$ is positive for small positive r.

We claim that $w^{*\prime}(r) \ge 0$ for $r > 0$. To verify this claim, we note that if $w^{*\prime}(r) < 0$ for some r, then $w^*(r)$ must have a local maximum. In that case, because solutions depend continuously on initial data, $w(r,d)$ must also have a local maximum for $d \in D$ sufficiently close to d^*. But Lemma 3.6 establishes that for $d \in D$, $w'(r,d) > 0$ on $(0, R(d)]$. Thus $w^*(r)$ cannot have a local maximum, hence $w^{*\prime}(r) \ge 0$ for all $r > 0$.

Lemma 3.3 then establishes that $w^{*\prime}(r) > 0$ for all $r > 0$. Since $w^*(r)$ is positive for small positive r, it immediately follows that $w^*(r) > 0$ for all $r > 0$.

It remains to show that $\lim_{r\to\infty} w^*(r) = \delta$. Since $w^*(r)$ is monotonically increasing and bounded above by δ, the limit $L \equiv \lim_{r\to\infty} w^*(r)$ exists, and $L \le \delta$.

We claim that $f(L) = 0$. To verify this claim, note that (1.15) yields

$$\frac{1}{2} w^{*\prime 2} + F(w^*) \le \frac{\ell(\ell+N-2)}{2} \frac{w^{*2}}{r^2} + \frac{2}{r^2} \int_0^r sF(w^*)\, ds.$$

Since $|w^*(r)| \le \delta$ we have

$$\frac{2}{r^2} \int_0^r sF(w^*)\, ds \le F(\delta).$$

Also, $F(w)$ is bounded below by F_0. Hence,

$$\frac{1}{2}w^{*\prime 2} < \frac{\ell(\ell+N-2)}{2}\frac{\delta^2}{r^2} + F(\delta) - F_0.$$

Since (from (1.14)) $w^{*\prime}(r)$ has a finite limit as $r \to 0$, we thus conclude that $|w^{*\prime}|$ is bounded. Taking the limit $r \to \infty$ in (1.13) then yields $\lim_{r\to\infty} w^{*\prime\prime}(r) = -f(L)$. Since $|w^{*\prime}|$ is bounded, the limiting value of $w^{*\prime\prime}$ must be zero. Therefore $f(L) = 0$, as claimed.

We have thus established that w^* limits to a positive zero L of f. It remains to show that $L = \delta$. Recall that $f(\beta) = f(\delta) = 0$ and $f(s) > 0$ on (β, δ). We will show for $d \in D$ that $w(r, d)$ reaches $\frac{\gamma+\beta}{2}$ at $r = b_d$ where b_d can be bounded independent of d. Given that fact, and the fact that the monotonic function w^* is approximated arbitrarily closely on any compact interval by a solution $w(r, d)$ with $d \in D$, it follows that $L \geq \frac{\gamma+\beta}{2} > \beta$, from which we conclude (since δ is the only zero of f greater than γ) that $L = \delta$.

To show that b_d is bounded, consider $d \in D$. Then w achieves δ. Thus, w reaches γ at some finite value, $r = c_d$, and since $w' > 0$ (by Lemma 3.6) we have $0 < w < \gamma$ on $(0, c_d)$. Using identity (1.15) at c_d gives

$$0 \leq \frac{1}{2}c_d^2 w'(c_d)^2 \leq \frac{\ell(\ell+N-2)}{2}\gamma^2 + 2\int_0^{c_d} rF(w(r))\,dr. \tag{4.12}$$

We now estimate the integral term on (a_d, b_d) and obtain (since $F \leq 0$ on $(0, c_d)$ and $0 < a_d < b_d < c_d$)

$$2\int_0^{c_d} rF(w(r))\,dr \leq 2\int_{a_d}^{b_d} rF(w(r))\,dr.$$

Next, on (a_d, b_d) we have $\alpha < w(r) < \frac{\beta+\gamma}{2}$. In particular, since $F < 0$ on $[\alpha, \frac{\beta+\gamma}{2}]$, there exists $C^2 > 0$ such that $F(s) \leq -C^2$ on $[\alpha, \frac{\beta+\gamma}{2}]$. Therefore,

$$2\int_{a_d}^{b_d} rF(w(r))\,dr \leq -C^2(b_d^2 - a_d^2).$$

Thus, (4.12) can be rewritten as

$$b_d^2 - a_d^2 \leq \frac{\ell(\ell+N-2)\gamma^2}{2C^2}.$$

Now recall that we proved in Lemma 4.3 that

$$a_d^\ell \leq \frac{\alpha}{d}.$$

Since $d \in D$, we have that $d > d^*$. Thus,

$$b_d^2 \leq \Big(\frac{\alpha}{d^*}\Big)^{2/\ell} + \frac{\ell(\ell+N-2)\gamma^2}{2C}.$$

Therefore,

$$b_d^2 \leq K,$$

where K is independent of $d \in D$, as claimed. This completes the proof of Lemma 4.7.

We now let B be the set of d-values for which the solution approaches the hilltop asymptotically. That is, we define

$$B = \{d > 0 \mid w(r,d) > 0 \text{ for } r > 0, \quad w'(r,d) > 0 \text{ for } r > 0, \text{ and } \lim_{r\to\infty} w(r,d) = \delta\}.$$

We know that $d^* \in B$ so that B is nonempty. Also, Lemma 4.1 shows that B is bounded away from zero. Now let

$$d_* = \inf B.$$

Remark. We have $0 < d_* \le d^*$. We expect in general that $B = \{d^*\}$, so that $d_* = d^*$, but we have not proven this assertion.

We now show that $d_* \in B$.

LEMMA 4.8. *$w(r,d_*) > 0$ for $r > 0$, $w'(r,d_*) > 0$ for $r > 0$ and $\lim_{r\to\infty} w(r,d_*) = \delta$.*

Proof. If $d_* \in B$, then the assertion of the lemma is trivially true. So suppose $d_* \notin B$. Then $d_* < d^*$. Consider a sequence $\{d_j \in B\}$ with $d_* < d_j < d^*$ such that $d_j \to d_*^+$. Because solutions depend continuously on initial conditions, the functions $w(r,d_j)$ converge uniformly on compact intervals to $w(r,d_*)$. Because each $d_j \in B$, it follows that $0 \le w(r,d_*) \le \delta$ and $w'(r,d_*) \ge 0$. Since $d_* > 0$, $w(r,d_*)$ is not identically zero. So, by Lemma 3.3, $w'(r,d_*) > 0$, hence $w(r,d_*) > 0$ for $r > 0$. Exactly as in the proof of Lemma 4.7, it can now be shown that $\lim_{r\to\infty} w(r,d_*) = \delta$. This completes the proof of Lemma 4.8.

LEMMA 4.9. *If $d < d_*$ and d is sufficiently close to d_*, then $w(r,d)$ has a local maximum.*

Proof. Since $0 < d < d_* \le d^*$, we have that $d \notin D$ and therefore $R(d) = \infty$. Now suppose $w(r,d)$ does not have a local maximum for all d sufficiently close to d_*. Then there is a sequence $\{d_j\}$ with $0 < d_j < d_*$ and $d_j \to d_*^-$ such that $w'(r,d_j) \ge 0$ on $[0,\infty)$. By Lemma 3.3, $w'(r,d_j) > 0$ on $(0,\infty)$ and thus $0 < w(r,d_j) < \delta$ on $(0,\infty)$. As in the proof of Lemma 4.7, $\lim_{r\to\infty} w(r,d_j) = L_j$, where $f(L_j) = 0$. But also by continuity of solutions of initial value problems, $w(r,d_j)$ converges uniformly on compact subsets to $w(r,d_*)$ as $j \to \infty$. Now since $w(r,d_*)$ limits to δ as $r \to \infty$, we can make L_j larger than β for d_j close enough to d_*. Since $f(L_j) = 0$, we must then have $L_j = \delta$. Hence, $d_j \in B$ for all sufficiently large j. This contradicts the hypothesis that $d_j < d_*$. Thus $w(r,d)$ must have a local maximum for all d sufficiently close to d_*. This proves Lemma 4.9.

As a consequence of Lemma 4.9, for d sufficiently close to d_*, the solution $w(r,d)$ of (1.13)–(1.14) has a first local maximum at $r = M_d$. We now show that this first local maximum occurs later and grows in amplitude as d approaches d_* from below.

LEMMA 4.10. *$\lim_{d\to d_*^-} M_d = \infty$ and $\lim_{d\to d_*^-} w(M_d,d) = \delta$.*

Proof. Lemma 4.9 establishes the existence of a first local maximum M_d for $d < d_*$ and d close to d_*. To prove that $M_d \to \infty$, suppose not. Then $M_d \le C < \infty$ and so there is a subsequence of the ds (again labeled d) and an M such that $M_d \to M$. By Lemma 4.5 we have then that $z(r,d) \to z(r,d_*)$ and $z'(r,d) \to z'(r,d_*)$ uniformly on $[0, M+1]$, and hence $w(r,d) \to w(r,d_*)$ and $w'(r,d) \to w'(r,d_*)$ uniformly on $[0, M+1]$. Since M_d is a local maximum for $w(r,d)$, we have that $w'(M_d,d) = 0$. On the other hand, $0 = w'(M_d,d) \to w'(M,d_*) > 0$ where the last inequality is from Lemma 4.8. This contradiction shows that $M_d \to \infty$ as $d \to d_*$.

To prove the second assertion of the lemma, consider $0 < d < d_*$ with d sufficiently close to d_* that $w(r,d)$ has a first local maximum at $r = M_d$. Because $w(r,d) > 0$ for small r, it follows that $0 < w(r,d) < w(M_d,d) < \delta$ for $r \in (0, M_d)$.

Since (by Lemma 4.8) $w(r,d_*) \to \delta$ as $r \to \infty$, given $\epsilon > 0$ there is r_ϵ such that $w(r,d_*) \geq \delta - \frac{\epsilon}{2}$ for all $r \geq r_\epsilon$. Because $w(r,d) \to w(r,d_*)$ uniformly on $[0,r_\epsilon]$ as $d \to d_*^-$, there is a number $d_\epsilon < d_*$ such that $|w(r,d) - w(r,d_*)| < \frac{\epsilon}{2}$ for all $r \in [0,r_\epsilon]$ whenever $d_\epsilon < d < d_*$. Because $M_d \to \infty$ as $d \to d_*^-$, we may furthermore choose d_ϵ such that additionally $M_d > r_\epsilon$ whenever $d_\epsilon < d < d_*$. Thus $\delta - \epsilon < w(r_\epsilon, d) < w(M_d, d) < \delta$ for $d_\epsilon < d < d_*$. Since ϵ may be chosen arbitrarily small, it follows that $w(M_d,d) \to \delta$ as $d \to d_*^-$, as claimed. This completes the proof of Lemma 4.10.

5. Energy analysis. We will now establish that there are solutions of the initial value problem (1.13)–(1.14) with arbitrarily many zeros. We will then finally prove the main theorem.

LEMMA 5.1. *Suppose that a solution $w(r)$ to (1.13) has successive local extrema $w_1 \equiv w(r_1)$ and $w_2 \equiv w(r_2)$, with $0 < r_1 < r_2 < \infty$ (and $w'(r) \neq 0$ for all $r \in (r_1, r_2)$.) Then $(w_1 - w_2)(f(w_1) - f(w_2)) > 0$. Thus f must be increasing somewhere in the interval between w_1 and w_2.*

Proof. First suppose $w_1 = w(r_1)$ is a local maximum of w. Then $w_2 = w(r_2)$ is a local minimum, and we have $w'(r_1) = 0$, $w''(r_1) \leq 0$, $w'(r_2) = 0$, $w''(r_2) \geq 0$, and $w_2 < w_1$. Evaluating (1.13) at r_1 and r_2 and subtracting, we obtain

$$(w''(r_2) - w''(r_1)) + \ell(\ell+N-2)\left[\frac{w_1}{r_1^2} - \frac{w_2}{r_2^2}\right] = f(w_1) - f(w_2). \tag{5.1}$$

Because $\frac{1}{r_1^2} > \frac{1}{r_2^2}$ and $w_1 > w_2$, the term in (5.1) in square brackets is strictly positive. Because $(w''(r_2) - w''(r_1)) \geq 0$, we see that $f(w_1) - f(w_2) > 0$, as claimed. This establishes the assertion of the lemma in the case when w_1 is a local maximum.

Next suppose $w_1 = w(r_1)$ is a local minimum of w. Then $w_2 = w(r_2)$ is a local maximum, and $w_2 > w_1$. Because f is odd, the function $\tilde{w}(r) \equiv -w(r)$ is also a solution to (1.13), and we note that $\tilde{w}_1 \equiv \tilde{w}(r_1) = -w_1$ is a local maximum of $\tilde{w}$, and that $\tilde{w}_2 \equiv \tilde{w}(r_2) = -w_2$ is a local minimum of $\tilde{w}$. Thus the reasoning in the first part of our proof applies to $\tilde{w}$, and we conclude that $f(\tilde{w}_1) - f(\tilde{w}_2) > 0$. Thus (again because f is odd) $f(w_1) - f(w_2) < 0$, while $w_1 - w_2 < 0$ also. This completes the proof of Lemma 5.1.

LEMMA 5.2. *If $w(r) = w(r,d)$ is a solution to (1.13)–(1.14) with $0 < d < d_*$, then*

$$|w'| \leq \sqrt{\ell(\ell+N-2)\frac{\delta^2}{r^2} + 2\,(F(\delta) - F_0)} \tag{5.2}$$

for all $r > 0$.

Proof. Identity (1.15) with the initial condition (1.14) yields

$$\tfrac{1}{2}w'^2 + F(w) \leq \frac{\ell(\ell+N-2)}{2}\frac{w^2}{r^2} + \frac{2}{r^2}\int_0^r s\,F(w(s))\,ds. \tag{5.3}$$

Since we know that $|w(r)| < \delta$ for all $r \geq 0$ and since $\max_{y\in[-\delta,\delta]} F(y) = F(\delta)$, we have $\int_0^r s\,F(w(s))\,ds \leq \frac{1}{2}r^2F(\delta)$. Hence (5.3) yields

$$\tfrac{1}{2}w'^2 \leq \frac{\ell(\ell+N-2)}{2}\frac{w^2}{r^2} + F(\delta) - F(w) < \frac{\ell(\ell+N-2)}{2}\frac{\delta^2}{r^2} + F(\delta) - F_0, \tag{5.4}$$

that is,

$$|w'| \le \sqrt{\ell(\ell+N-2)\frac{\delta^2}{r^2} + 2\,(F(\delta) - F_0)},$$

as asserted.

For a solution $w(r)$ to (1.13), we define the "energy"

$$E(r) \equiv \frac{1}{2}w'(r)^2 + F(w(r)). \tag{5.5}$$

Differentiating E and using (1.13), we find

$$E' = -\frac{N-1}{r}w'^2 + \frac{\ell(\ell+N-2)}{r^2}ww'. \tag{5.6}$$

We now estimate the loss of energy on an interval of monotonicity for the solution. Lemma 5.3 establishes that the energy loss is bounded independent of the length of the interval.

LEMMA 5.3. *Suppose $w(r) = w(r,d)$ is a solution to* (1.13)–(1.14) *with $0 < d < d_*$, such that $w'(r) \neq 0$ for all $r \in (r_1, r_2)$, where $0 < r_1 < r_2$. Then*

$$E(r_1) - E(r_2) \le \frac{C_1}{r_1} + \frac{C_2}{r_1^2}, \tag{5.7}$$

where C_1 and C_2 are positive constants that are independent of d, r_1, and r_2.

Proof. Since $w'(r) \neq 0$ for all $r \in (r_1, r_2)$, it follows that w' has a constant sign on (r_1, r_2), and that w can change sign at most once on (r_1, r_2). Thus ww' can change sign at most once on (r_1, r_2). Let (s_1, s_2) be the maximal subinterval of (r_1, r_2) on which $ww' < 0$. (The interval (s_1, s_2) could be empty, in which case the corresponding integrals below are taken to be zero.) Thus $0 < r_1 \le s_1 \le s_2 \le r_2$.

We now estimate $E(r_1) - E(r_2)$.

$$\begin{aligned}
E(r_1) - E(r_2) &= -\int_{r_1}^{r_2} E'(r)\, dr \\
&= \int_{r_1}^{r_2} \frac{N-1}{r}w'^2\, dr - \int_{r_1}^{r_2} \frac{\ell(\ell+N-2)}{r^2}\, ww'\, dr \\
&\le \int_{r_1}^{r_2} \frac{N-1}{r}w'^2\, dr - \int_{s_1}^{s_2} \frac{\ell(\ell+N-2)}{r^2}\, ww'\, dr \\
&\le \frac{N-1}{r_1}\int_{r_1}^{r_2} w'^2\, dr + \frac{\ell(\ell+N-2)}{r_1^2}\int_{s_1}^{s_2} (-ww')\, dr \\
&= \frac{N-1}{r_1}\int_{r_1}^{r_2} w'^2\, dr + \frac{\ell(\ell+N-2)}{2r_1^2}\left[w(s_1)^2 - w(s_2)^2\right] \\
&\le \frac{N-1}{r_1}\int_{r_1}^{r_2} w'^2\, dr + \frac{\ell(\ell+N-2)}{2}\frac{\delta^2}{r_1^2}.
\end{aligned} \tag{5.8}$$

To estimate the remaining integral we note that for $r \in (r_1, r_2)$, Lemma 5.2 yields

$$|w'(r)| \le \sqrt{\ell(\ell+N-2)\frac{\delta^2}{r_1^2} + 2\,(F(\delta) - F_0)}. \tag{5.9}$$

Thus

$$\int_{r_1}^{r_2} w'^2\, dr = \int_{r_1}^{r_2} |w'(r)|\ |w'(r)|\ dr \leq \sqrt{\ell(\ell+N-2)\delta^2 r_1^{-2} + 2F(\delta) - 2F_0}\ \int_{r_1}^{r_2} |w'(r)|\ dr.$$

Since $w'(r) \neq 0$ for $r \in (r_1, r_2)$, we have

$$\begin{aligned} \int_{r_1}^{r_2} w'^2\, dr &\leq \sqrt{\ell(\ell+N-2)\delta^2 r_1^{-2} + 2F(\delta) - 2F_0}\ |w(r_2) - w(r_1)| \\ &\leq 2\delta\sqrt{\ell(\ell+N-2)\delta^2 r_1^{-2} + 2F(\delta) - 2F_0} \\ &\leq 2\sqrt{2}\delta\sqrt{F(\delta) - F_0} + 2\delta^2 r_1^{-1}\sqrt{\ell(\ell+N-2)}. \end{aligned} \tag{5.10}$$

Substituting (5.10) into (5.8), we obtain (5.7), as claimed, with

$$C_1 = C_1(N, \ell, f) = 2\sqrt{2}(N-1)\delta\sqrt{F(\delta) - F_0}$$

and

$$C_2 = C_2(N, \ell, f) = \left(\tfrac{1}{2}\ell(\ell+N-2) + 2(N-1)\sqrt{\ell(\ell+N-2)}\right)\delta^2.$$

This completes the proof of Lemma 5.3.

For brevity, we define $h(r) \equiv \frac{C_1}{r} + \frac{C_2}{r^2}$, where C_1 and C_2 are the constants of Lemma 5.3.

LEMMA 5.4. *Suppose $w(r) = w(r, d)$ is a solution to* (1.13)–(1.14) *with $0 < d < d_*$. Suppose that $|w|$ has a local maximum at $r = p$, with $F(w(p)) - h(p) > 0$. Then w' has a zero for some value of r larger than p.*

Proof. Suppose by way of contradiction that $w' \neq 0$ for all $r > p$. Then, because $|w|$ is bounded, w has a limit $L \equiv \lim_{r\to\infty} w(r)$. Using the fact that $|w'|$ is bounded and taking limits in (1.13) shows that $\lim_{r\to\infty} w''(r) = -f(L)$. As earlier, this limiting value of w'' must vanish because $|w'|$ is bounded. Thus $f(L) = 0$ and $w'(r)$ has a limit as $r \to \infty$ which must then vanish because $|w|$ is bounded. It then follows that $\lim_{r\to\infty} E(r) = F(L)$.

Now from Lemma 5.3 we have $E(r) \geq E(p) - h(p) = F(w(p)) - h(p) > 0$ for all $r > p$. Thus $F(L) > 0$, hence $\gamma < |L| \leq \delta$. Since the only zeros of f with magnitude between γ and δ are $\pm\delta$, we find that $\lim_{r\to\infty} |w(r)| = \delta$. But Lemma 3.1 asserts that $|w(r)| \leq |w(p)| < \delta$ for all $r \geq p$. Thus $|w|$ cannot have limit δ. This contradiction establishes that w' must have a zero at $r > p$, and completes the proof of Lemma 5.4.

We next establish that there are solutions to the initial value problem (1.13)–(1.14) with arbitrarily many zeros. This result, in Lemma 5.5, elucidates the solution set structure as a function of d and is an essential part of the proof of the main theorem. Our proof of Lemma 5.5 is based on the fact, established by Lemma 5.3, that a solution's loss of energy on an interval of monotonicity is bounded independent of the length of the interval.

LEMMA 5.5. *Given any positive integer n, there is a number e_n between 0 and d_* such that the solution of* (1.13)–(1.14) *has at least n positive zeros for all $d \in (e_n, d_*)$.*

Proof. By hypothesis there is a number $\lambda \in (\gamma, \delta)$ such that f is strictly decreasing on $[\lambda, \delta]$. Recall that F is then positive, increasing, and concave downward on $[\lambda, \delta]$.

By Lemmas 4.9 and 4.10, $w(r,d)$ has a first local maximum $r = M_d$ for d sufficiently close to d_*, and $\lim_{d\to d_*^-} M_d = \infty$ and $\lim_{d\to d_*^-} w(M_d,d) = \delta$. Let $c \in (0,d_*)$ be such that $\lambda < w(M_d,d) < \delta$ for all $d \in (c,d_*)$.

Because $F(\delta) > F(\lambda)$, given n we may choose $e_n \in (c,d_*)$ so close to d_* that $F(w(M_d,d)) - nh(M_d) > F(\lambda)$ for all $d \in (e_n,d_*)$. Consider fixed $d \in (e_n,d_*)$. By Lemma 5.4, w' has a next zero $r = m_1 > M_d$. By Lemma 5.3, $E(m_1) \geq E(M_d) - h(M_d)$. Thus $F(w(m_1)) \geq F(w(M_d)) - h(M_d) > F(\lambda)$, and therefore $|w(m_1)| > \lambda$. Because f is strictly decreasing on $[\lambda,\delta]$ and because $w(M_d) \in (\lambda,\delta)$, by Lemma 5.1 it cannot be that $w(m_1) \in (\lambda,\delta)$. Thus $w(m_1) \in (-\delta,-\lambda)$. Hence w has a zero between $r = M_d$ and $r = m_1$.

If $n > 1$, we may now repeat the argument: Because

$$\begin{aligned} F(w(m_1)) - h(m_1) &> F(w(m_1)) - h(M_d) \\ &\geq F(w(M_d)) - 2h(M_d) \\ &> F(\lambda), \end{aligned}$$

by Lemma 5.4, w' has a next zero $r = m_2 > m_1$. By Lemma 5.3, $E(m_2) \geq E(m_1) - h(m_1)$. Thus $F(w(m_2)) \geq F(w(m_1)) - h(m_1) > F(\lambda)$, and hence $|w(m_2)| > \lambda$. Because f is strictly decreasing on $(-\delta,-\lambda)$ and because $w(m_1) \in (-\delta,-\lambda)$, by Lemma 5.1 we have $w(m_2) \in (\lambda,\delta)$. Hence w has another zero between $r = m_1$ and $r = m_2$.

We may continue in this fashion to show that there are n positive zeros of w. This concludes the proof of Lemma 5.5.

The last major result required for the proof of the main theorem is the fact that as d is varied new zeros in the solution to (1.13)–(1.14) appear one at a time, and changes in the number of zeros occur at isolated values of d. This result is established in Lemma 5.9. The following three lemmas are used in the proof of Lemma 5.9.

LEMMA 5.6. *Suppose w is a nontrivial solution of* (1.13) *with $w(t) = 0$ for some $t > 0$, and suppose $p < t$ is such that $w'(p) = 0$ and $w'(r) \neq 0$ for $r \in (p,t)$. Then $|w(p)| > \gamma$.*

Proof. Because $w'(r) \neq 0$ for $r \in (p,t)$, we have $w(r)w'(r) < 0$ for $r \in (p,t)$. So by (5.6), $E'(r) < 0$ for $r \in (p,t)$. Thus $E(p) > E(t) \geq 0$, hence $F(w(p)) = E(p) > 0$, hence $|w(p)| > \gamma$, as asserted.

LEMMA 5.7. *If w is a nontrivial solution to* (1.13) *such that $\lim_{r\to\infty} w(r) = 0$, then $w'(r) \neq 0$ for all sufficiently large r.*

Proof. Suppose on the contrary that w is nontrivial and there is a sequence $p_j \to \infty$ such that $w(p_j) \to 0$ as $j \to \infty$ and $w'(p_j) = 0$ for all j.

We note first that if $w''(p) = w'(p) = 0$ for some $p > 0$, then, by (1.13), either $w(p) = 0$, or $w(p) \neq 0$ and $f(w(p))$ has the same sign as w. Since w is nontrivial, by uniqueness of solutions to initial value problems we cannot have $w(p) = w'(p) = 0$, so it must be that $w(p)$ and $f(w(p))$ have the same sign. Thus $|w(p)| > \alpha$. Since $w(p_j) \to 0$ as $j \to \infty$, it must be that $w''(p_j) \neq 0$ for sufficiently large j.

Under our supposition, it follows that there is an increasing sequence of local maxima M_j of $|w|$, with $M_j \to \infty$ and $0 \neq w(M_j) \to 0$ as $j \to \infty$. Because $F(y) < 0$ for $|y| \in (0,\gamma)$, we have $F(w(M_J)) < 0$ for some sufficiently large J. Thus by Lemma 3.1, $F(w(M_j)) \leq F(w(M_J)) < 0$ for all $j > J$. This contradicts the fact that $\lim_{j\to\infty} F(w(M_j)) = F(0) = 0$. This contradiction shows that there cannot be a sequence p_j with the properties that we assumed, thereby proving the lemma.

LEMMA 5.8. *Suppose w is a nontrivial solution to* (1.13) *such that $\lim_{r\to\infty} w(r) = 0$. Suppose $w'(p) = 0$ and $w'(r) \neq 0$ for all $r > p$. Then $|w(p)| > \gamma$.*

Proof. Since $|w|$ is monotonically decreasing to 0 on (p,∞), as in Lemma 5.6 it follows that $E'(r) < 0$ for all $r > p$. Since $E(r)$ is bounded below (by F_0), $E(r) = \frac{1}{2}w'(r)^2 + F(w(r))$ has a limit as $r \to \infty$. Since $F(w(r)) \to 0$ as $r \to \infty$, it follows that $|w'|$ has a limit, which must vanish. Thus $F(w(p)) = E(p) > \lim_{r\to\infty} E(r) = 0$, hence $|w(p)| > \gamma$, as asserted.

LEMMA 5.9. *Suppose $w(r) = w(r, d_c)$ is a solution to (1.13)–(1.14) with $d = d_c \in (0, d_*)$, such that $w(r)$ has exactly k zeros, and such that $\lim_{r\to\infty} w(r, d_c) = 0$. If d is sufficiently close to d_c then the solution $w(r, d)$ has at most $(k+1)$ zeros.*

Proof. We wish to show that for d near d_c, $w(\cdot, d)$ has at most $(k+1)$ zeros in $(0,\infty)$. So we suppose there is a sequence of values $d_j \in (0, d_*)$ converging to d_c as $j \to \infty$ such that $w(\cdot, d_j)$ has at least $(k+1)$ zeros in $(0,\infty)$. (If there is no such sequence then the lemma is proven.) We write $w_j(r) \equiv w(r, d_j)$ and we denote by r_j the $(k+1)$st zero of w_j, counting from the smallest. We set $w_0(r) \equiv w(r, d_c)$.

By Lemma 5.7, $|w_0(r)|$ decreases monotonically to zero for sufficiently large r. Since $|w_0(r)| < \delta$ for all r, there is a largest number p_0 for which $w_0'(p_0) = 0$. By Lemma 5.8, $|w_0(p_0)| > \gamma$. Let $q_0 > p_0$ be such that $|w_0(q_0)| = \gamma$. We note that q_0 is unique because $|w_0(r)|$ is monotonically decreasing for $r > p_0$.

Let $\epsilon_2 \equiv \min\left\{\frac{1}{2}\beta, \frac{1}{2}(|w_0(p_0)| - \gamma)\right\}$. Given $\epsilon_1 \in (0, \epsilon_2)$, let $q_0^+ > q_0$ be defined by $|w_0(q_0^+)| = \gamma - \epsilon_1$, and let $q_0^- \in (p_0, q_0)$ be defined by $|w_0(q_0^-)| = \gamma + \epsilon_1$. Note that q_0^+ and q_0^- approach q_0 as $\epsilon_1 \to 0^+$.

Since $w_0'(q_0) \neq 0$, there is a number $m_0 > 0$ such that $\min_{r\in[q_0^-, q_0^+]} |w_0'(r)| > m_0 > 0$ independent of $\epsilon \in (0, \epsilon_1)$. Let $\epsilon_0 \equiv \min\left\{\frac{1}{2}m_0, \epsilon_1\right\}$. Let $L_0 > q_0$ be the unique number such that $|w_0(L_0)| = \frac{1}{2}\beta$.

Because solutions depend continuously on initial data, we know that $w_j \to w_0$ and $w_j' \to w_0'$ uniformly on compact sets as $j \to \infty$. Thus, given $\epsilon \in (0, \epsilon_0)$ and $L > L_0$, there is a number $J(\epsilon, L)$ so large that for all $j > J(\epsilon, L)$ we have $\sup_{r\in[0,L]} |w_j(r) - w_0(r)| < \epsilon$ and $\sup_{r\in[0,L]} |w_j'(r) - w_0'(r)| < \epsilon$. In particular,

$$|w_j(q_0^+)| \le |w_0(q_0^+)| + |w_j(q_0^+) - w_0(q_0^+)| < \gamma - \epsilon_1 + \epsilon < \gamma$$

and

$$|w_j(q_0^-)| \ge \big||w_0(q_0^-)| - |w_j(q_0^-) - w_0(q_0^-)|\big| > \gamma + \epsilon_1 - \epsilon > \gamma$$

and, for all $r \in [q_0^-, q_0^+]$,

$$\begin{aligned} |w_j'(r)| &= \big|w_0'(r) + (w_j'(r) - w_0'(r))\big| \\ &\ge \big||w_0'(r)| - |w_j'(r) - w_0'(r)|\big| \\ &> m_0 - \epsilon > \tfrac{1}{2}m_0 > 0. \end{aligned}$$

Therefore, $|w_j(q_j)| = \gamma$ for a unique $q_j \in (q_0^-, q_0^+)$. We note that because ϵ_1 may be chosen arbitrarily small, we have $q_j \to q_0$ as $j \to \infty$.

For $j > J(\epsilon, L)$ it also follows that $|w_j(r)| < \gamma$ for all $r \in (q_j, L]$. Furthermore, $|w_j(L)| < \beta$ for all $j > J(\epsilon, L)$.

CLAIM. *For all sufficiently large j, $|w_j(r)| < \gamma$ for all $r > q_j$.*

Proof of Claim. Fix $\epsilon \in (0, \epsilon_0)$ and $L > L_0$. We already know that for $j > J(\epsilon, L)$ we have $|w_j(r)| < \gamma$ for $r \in (q_j, L]$. So suppose by way of contradiction that there exists a subsequence $d_j \to d_c$ (again labeled by j) such that, for all $j > J(\epsilon, L)$, $|w_j(Q_j)| = \gamma$ for some smallest $Q_j > L$.

By Lemma 3.1, $|w_j|$ cannot have a local maximum at any point in (q_j, Q_j). It follows, because $|w_j(L)| < \beta$, that there are numbers s_j and t_j with $L < s_j < t_j < Q_j$ such that $|w_j(s_j)| = \beta$ and $|w_j(t_j)| = \frac{1}{2}(\beta + \gamma) > \beta$ and $|w_j|$ is increasing on (s_j, t_j).

We now apply (1.15) between q_j and Q_j to obtain

$$\frac{1}{2}Q_j^2 w'(Q_j)^2 + (N-2)\int_{q_j}^{Q_j} s w_j'(s)^2 ds = \frac{1}{2}q_j^2 w'(q_j)^2 + 2\int_{q_j}^{Q_j} sF(w(s))ds. \tag{5.11}$$

We now estimate the integral term on the right side of (5.11). Because $|w(s)| < \gamma$ for $s \in (q_j, Q_j)$, we have $F(w(s)) \le 0$ for $s \in (q_j, Q_j)$. Thus

$$\begin{aligned} 2\int_{q_j}^{Q_j} sF(w(s))ds &\le 2\int_{s_j}^{t_j} sF(w(s))ds \le 2F(\tfrac{1}{2}(\beta+\gamma))\int_{s_j}^{t_j} s\,ds \\ &= F(\tfrac{1}{2}(\beta+\gamma))(t_j - s_j)(t_j + s_j). \end{aligned} \tag{5.12}$$

Now, Lemma 5.2 establishes that $|w_j'(r)|$ is bounded for $r \ge p_0$ by

$$w'_{\max} \equiv \sqrt{\ell(\ell+N-2)\frac{\delta^2}{p_0^2} + 2\,(F(\delta) - F_0)}$$

independent of j. We thus have

$$\tfrac{1}{2}(\beta+\gamma) - \beta = |w_j(t_j) - w_j(s_j)| \le w'_{\max}(t_j - s_j),$$

that is, $t_j - s_j \ge \frac{(\gamma-\beta)}{2w'_{\max}}$. Thus, since $F(\frac{1}{2}(\beta+\gamma)) < 0$ and $L < s_j < t_j$, (5.12) yields

$$2\int_{q_j}^{Q_j} sF(w(s))ds \le \frac{(\gamma-\beta)F(\frac{1}{2}(\beta+\gamma))}{w'_{\max}} L. \tag{5.13}$$

Thus by choosing $L > L_0$ sufficiently large we may make the term $2\int_{q_j}^{Q_j} sF(w(s))ds$ in (5.11) negative and arbitrarily large in magnitude. Since $q_j \to q_0$ as $j \to \infty$, the term $\frac{1}{2}q_j^2 w'(q_j)^2$ on the right side of (5.11) is bounded independent of j. Thus by choosing L sufficiently large, we can make the right side of (5.11) negative. Since the left side of (5.11) is manifestly positive, we arrive at a contradiction, thus establishing that, for sufficiently large j, $|w_j(r)| < \gamma$ for all $r > q_j$, as claimed.

To complete the proof of Lemma 5.9, we note that if w_j has a zero z_j beyond r_j, then $|w_j|$ has a local maximum at some point p between r_j and z_j. By Lemma 5.6, this local maximum must occur at an amplitude $|w_j(p)| > \gamma$, which the claim shows is impossible for sufficiently large j. Therefore, for sufficiently large j, $w_j(r)$ has at most the single zero r_j for $r > q_j$. This completes the proof of Lemma 5.9.

Proof of the Main Theorem. To prove the main theorem, we follow exactly the same steps as in [7, section 5]. The same proof technique applies despite the different hypotheses on the large-amplitude behavior of f in [7] because the large-r behavior of the solutions under study is governed by the small-amplitude features of f, which are exactly the same here as in [7]. The only difference in the proof is that here we use identity (1.15), which is a generalization of the Pohozaev identity used in [7].

In particular, we define

$$A_0 = \{d \in (0, d_*) \,|\, w(r,d) > 0 \text{ for all } r > 0\}.$$

By Lemma 4.1, A_0 is nonempty. Also, A_0 is bounded above by d_*. Setting

$$d_0 = \sup A_0,$$

we know from Lemma 5.5 that $d_0 < d_*$. As in [7, section 5], we can show that $w(r, d_0) > 0$ for $r > 0$, $w'(r, d_0) < 0$ for large r, and $\lim_{r\to\infty} w(r, d_0) = 0$. Next, we define

$$A_1 = \{d \in (d_0, d_*) \,|\, w(r, d) \text{ has exactly one zero in } (0, \infty)\}.$$

First, if $d_0 < d < d_*$, then $w(r, d)$ must have at least one zero in $(0, \infty)$ (by definition of d_0). Also, by Lemma 5.9, $w(r, d)$ has at most one zero in $(0, \infty)$ if d is close enough to d_0. Thus, A_1 is nonempty, and clearly A_1 is bounded above by d_*. Setting

$$d_1 = \sup A_1,$$

we know from Lemma 5.5 that $d_1 < d_*$. And again as in [7, section 5], we can show that $w(r, d_1)$ has exactly one zero and vanishes as $r \to \infty$. Continuing by induction, the main theorem is proved.

REFERENCES

[1] H. Berestycki and P. L. Lions, *Nonlinear scalar field equations*, I *and* II, Arch. Rational Mech. Anal., 82 (1983), pp. 313–375.

[2] M. Berger, *Nonlinearity and Functional Analysis*, Academic Press, New York, 1977.

[3] E. Deumens and H. Warchall, *Explicit construction of all spherically symmetric solitary waves for a nonlinear wave equation in multiple dimensions*, Nonlinear Anal., 12 (1988), pp. 419–447.

[4] G. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1995.

[5] B. Gidas, W.-M. Ni, and L. Nirenberg, *Symmetry and related properties via the maximum principle*, Comm. Math. Phys., 68 (1979), pp. 209–243.

[6] M. Grillakis, *Existence of nodal solutions of semilinear equations in* $\mathbb{R}^N$, J. Differential Equations, 85 (1990), pp. 367–400.

[7] J. Iaia and H. Warchall, *Nonradial solutions of a semilinear elliptic equation in two dimensions*, J. Differential Equations, 119 (1995), pp. 533–558.

[8] J. Iaia, H. Warchall, and F. B. Weissler, *Localized solutions of sublinear elliptic equations: Loitering at the hilltop*, Rocky Mountain J. Math., 27 (1997), pp. 1131–1157.

[9] C. Jones and T. Küpper, *On the infinitely many solutions of a semilinear elliptic equation*, SIAM J. Math. Anal., 17 (1986), pp. 803–835.

[10] G. King, *Explicit Multidimensional Solitary Waves*, Master's thesis, University of North Texas, Denton, TX, 1990.

[11] P. L. Lions, *Solutions complexes d'équations elliptiques semilinéaires dans* $\mathbb{R}^N$, C. R. Acad. Sci. Paris Sér. I Math., 302 (1986), pp. 673–676.

[12] K. McLeod and J. Serrin, *Uniqueness of positive radial solutions of* $\Delta u + f(u) = 0$ *in* $\mathbb{R}^N$, Arch. Rational Mech. Anal., 99 (1987), pp. 115–145.

[13] K. McLeod, W. C. Troy, and F. B. Weissler, *Radial solutions of* $\Delta u + f(u) = 0$ *with prescribed number of zeros*, J. Differential Equations, 83 (1990), pp. 368–378.

[14] L. A. Peletier and J. Serrin, *Uniqueness of positive solutions of semilinear equations in* $\mathbb{R}^N$, Arch. Rational Mech. Anal., 81 (1983), pp. 181–197.

[15] W. Strauss, *Existence of solitary waves in higher dimensions*, Comm. Math. Phys., 55 (1977), pp. 149–162.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 140–154

WEAK ASYMPTOTIC STABILITY OF SECOND-ORDER EVOLUTION EQUATIONS BY NONLINEAR AND NONMONOTONE FEEDBACKS*

JUDITH VANCOSTENOBLE†

Abstract. We consider the problem of asymptotic decay as $t \to +\infty$ of solutions of an abstract evolution equation of second order with a nonlinear and nonmonotone feedback. Weak asymptotic stability of the global solutions is proved. This abstract result can be applied to different types of equations (wave, beam, and plate equations) and to different types of controls (interior, boundary, or pointwise controls). In particular, we significantly improve several earlier results on the asymptotic stability of the wave equation in a bounded domain with an interior or boundary control.

Key words. nonlinear second-order PDE, asymptotic behavior, stabilization, nonmonotone feedback, distributed control, boundary control, pointwise control

AMS subject classifications. 35L70, 35B40

PII. S0036141098332378

1. Introduction.

We consider the problem of asymptotic decay as $t \to +\infty$ of solutions of an abstract evolution equation of second order with a nonlinear and nonmonotone feedback. More precisely, we consider some evolution equations of second order for which strong compactness of trajectories does not hold while weak compactness does. We give general conditions for which weak asymptotic stability as $t \to +\infty$ of global solutions will occur. As an illustration, let us consider the following classical control problem.

EXAMPLE 1.1. *Let Ω be a bounded open subset of $\mathbb{R}^N, N \geq 1$ with a smooth boundary Γ, and let (Γ_0, Γ^*) be a nontrivial partition of Γ (i.e., Γ_0, Γ^* are closed, $\mathrm{int}(\Gamma^*) \neq \emptyset$, $\mathrm{int}(\Gamma_0) \neq \emptyset$, and $\mathrm{int}(\Gamma^*) \cap \mathrm{int}(\Gamma_0) = \emptyset$). Let $q : \mathbb{R} \to \mathbb{R}$ be continuous and such that*

$$\forall \lambda \in \mathbb{R},\ \lambda q(\lambda) \geq 0 \text{ and } \forall \lambda > 0,\ q(\lambda) > 0. \tag{1.1}$$

Let $a \in L^\infty(\Gamma_0)$ such that $a(.) > 0$ on Γ_0. We set $V = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma^*\}$, and we consider the following equation:

$$\begin{cases} u_{tt} - \Delta u = 0 & \text{for } (t,x) \in \mathbb{R}_+ \times \Omega, \\ u = 0 & \text{for } (t,x) \in \mathbb{R}_+ \times \Gamma^*, \\ \frac{\partial u}{\partial \nu} = -a(x) q(u_t) & \text{for } (t,x) \in \mathbb{R}_+ \times \Gamma_0, \\ u(0) = u_0, u_t(0) = v_0 & \text{for } x \in \Omega, \end{cases} \tag{1.2}$$

where the initial data (u_0, v_0) are given in $V \times L^2(\Omega)$.

*Received by the editors January 9, 1998; accepted for publication January 23, 1998; published electronically October 20, 1998.
http://www.siam.org/journals/sima/30-1/33237.html

†Institut de Recherche Mathématique Avancée, Université Louis Pasteur et CNRS, 7, rue René Descartes, 67084 Strasbourg Cédex, France (vancoste@math.u-strasbg.fr). This work was done while the author was working for the Projet NUMATH, Institut Élie Cartan, Université Henri Poincaré Nancy 1, UMR CNRS 9973 et INRIA-Lorraine.

If q is monotone increasing and satisfies

$$|q(\lambda)| \leq A + B|\lambda|^r \begin{cases} \text{with } r \leq min(\frac{N}{N-2}, 2) & \text{if } N \geq 3, \\ \text{with } r \leq 2 & \text{if } N = 2, \\ \text{and no condition} & \text{if } N = 1, \end{cases} \tag{1.3}$$

then strong asymptotic stability of the solution is known (a proof and references can be found in [1]*). But if q does not verify the hypothesis* (1.3)*, the problem of strong asymptotic stability of the solution is still open even if q is monotone increasing. We will prove that in this case there is at least weak asymptotic stability of the solution. Actually if we suppress the hypothesis of monotonicity of q, and if we suppose only that $\forall \alpha > 0,\ \inf\{q(\lambda) \mid \lambda \geq \alpha\} > 0$, then we still obtain weak asymptotic stability of all global solutions.*

More generally, we will use an abstract framework and we will prove weak asymptotic stability of the global solutions of an abstract evolution equation of second order with a nonlinear and nonmonotone feedback. The use of this abstract framework underlines the essential properties of the considered equation which are necessary for the proof. Moreover, our abstract result of weak asymptotic stability can be applied to many equations (including, for example, wave- and platelike equations with interior feedbacks but also with boundary or pointwise controls). Our result improves significantly several earlier results on the subject (see, for example, [3] , [6] , [1] , [9]) insofar as we need neither hypothesis of monotonicity of the control nor condition restricting its asymptotic growth.

2. Results.

2.1. Abstract framework and theorem of weak asymptotic stability. We suppose the following.

• (H_1)(i): Let X be a locally compact space, and let μ be a positive measure such that $\mu(X) < +\infty$. We denote by $K(X)$ the space of continuous and compactly supported functions from X into $\mathbb{R}$, and we denote by H the Hilbert space $L^2(X, \mu)$ equipped with the scalar product

$$\forall u, v \in H,\ (u, v)_H = \int_X u(x)v(x)d\mu(x).$$

• (H_1)(ii): Let A be a linear operator on H with dense domain $D(A)$. We assume that A is self-adjoint and coercive and that the resolvent of A is compact. We define $V = D(A^{\frac{1}{2}})$ equipped with the scalar product

$$\forall u, v \in V,\ (u, v)_V = (A^{\frac{1}{2}}u, A^{\frac{1}{2}}v)_H = \langle \tilde{A}u, v\rangle_{V',V},$$

where $\tilde{A} \in \mathcal{L}(V; V')$ is defined by the bilinear form $(.,.)_V$ and extends A. As usual, we identify H with its dual. Then $V \hookrightarrow H \hookrightarrow V'$ with the following relation:

$$\forall h \in H, \forall v \in V,\ \langle h, v\rangle_{V',V} = (h, v)_H.$$

Moreover, we suppose that $\mathcal{E} = K(X) \cap V$ is dense in V.

• (H_1)(iii): Let Y be a subspace of X (i.e., $Y \subset X$), and let m be a positive measure on Y such that $m(Y) < +\infty$. For all $v \in K(X)$, we denote by $\tau(v)$ the restriction of v to Y and we suppose that

$$\exists C > 0/\ \forall v \in \mathcal{E},\ \|\tau(v)\|_{L^1(Y,m)} \leq C\|v\|_V.$$

Consequently, the linear mapping τ can be extended to a linear continuous application from $V \to L^1(Y,m)$. So we can define the restriction to Y of all elements of V. For all $v \in V$, we will simply denote by v this restriction (see the remark below for various examples).

• (H_2)(i): Let $q : \mathbb{R} \to \mathbb{R}$ be continuous and such that

$$\forall \lambda \in \mathbb{R},\ \lambda q(\lambda) \geq 0, \tag{2.1}$$

$$\forall \alpha > 0,\ \inf\{q(\lambda) \mid \lambda \geq \alpha\} > 0. \tag{2.2}$$

• (H_2)(ii): Let $a \in L^\infty(Y,m)$ with $a(y) \geq 0$ almost everywhere for measure m. We define $Q : D(Q) \to V'$ by

$$D(Q) = \Big\{ v \in V \mid \forall \varphi \in \mathcal{E},\ a(.)q(v)\varphi \in L^1(Y,m) \\ \text{and } \exists C_v > 0 \text{ such that } \left| \int_Y aq(v)\varphi dm \right| \leq C_v \|\varphi\|_V\ \forall \varphi \in \mathcal{E} \Big\} \tag{2.3}$$

and

$$\forall v \in D(Q),\ \forall \varphi \in \mathcal{E} \cup \{v\},\ \langle Q(v), \varphi \rangle_{V',V} = \int_Y a(y)q(v(y))\varphi(y)dm(y). \tag{2.4}$$

• (U): Moreover, for asymptotic stability we will need the following "uniqueness" property:

$$\big(\varphi \in V,\ A\varphi = \omega^2 \varphi \ \textbf{ and }\ a(y)\varphi(y) = 0 \text{ m-a.e. in } Y\big) \ \Rightarrow \varphi(x) = 0 \ \mu\text{-a.e. } x \in X.$$

We consider the following problem for (u_0, v_0) given in $V \times H$:

$$\begin{cases} u_{tt} + \tilde{A}u = -Q(u_t), & t \in \mathbb{R}_+, \\ u \in V, & t \in \mathbb{R}_+, \\ u(0) = u_0,\ u_t(0) = v_0, \end{cases} \tag{2.5}$$

and we obtain the following result of weak asymptotic stability.

THEOREM 2.1. *We assume (H_1), (H_2), and (U) hold. Let (u_0, v_0) be given in $V \times H$. We suppose that there exists $u(t; u_0, v_0)$, a solution of the problem* (2.5) *in the following sense:*

$$\begin{cases} (u, u_t) \in \mathcal{C}([0, +\infty[; V \times H), \\ u_t \in L^2_{loc}(]0, +\infty[; V), \\ u_{tt} \in L^2_{loc}(]0, +\infty[; H), \\ \tilde{A}u + Q(u_t) \in L^2_{loc}(]0, +\infty[; H), \end{cases} \tag{2.6}$$

and u satisfies the equation (2.5) *almost everywhere (a.e.) t in V'. Then $(u(t), u_t(t)) \rightharpoonup (0,0)$ weakly in $V \times H$ as $t \to +\infty$.*

Remark. The abstract framework can be applied to the different types of controls as follows:

- In the case of an equation in a bounded open Ω of $\mathbb{R}^N$, $N \geq 1$ with an interior control, we will set $X = Y = \Omega$ and $\mu = m = dx$ where dx is the Lebesgue's measure in $\mathbb{R}^N$.

- In the case of an equation in a bounded open Ω of $\mathbb{R}^N$, $N \geq 1$ with a boundary control on $\Gamma = \partial\Omega$, we will set $X = \overline{\Omega}$, $\mu = dx$, and $Y = \Gamma$, $m = d\nu$ where $d\nu$ is the superficial measure on Γ.
- In the case of an equation in a bounded open Ω of $\mathbb{R}^N$, $N \geq 1$ with a pointwise control at $p \in \Omega$, we will set $X = \Omega$, $\mu = dx$, $Y = \{p\}$, and m will be the Dirac masse $\delta(\cdot - p)$.
- More generally, we can choose $X = \Omega$, $\mu = dx$, and Y a set of positive "capacity" for the capacity associated with the norm of V and m carried by Y (see [1]).

Remark. Note that the hypothesis (U) is a *necessary condition* of weak asymptotic stability of all global solutions of (2.5). Indeed, if (U) is not satisfied then there exists $\varphi \in V$ and $\omega \in \mathbb{R}$ such that $A\varphi = \omega^2\varphi$ and $a(y)\varphi(y) = 0$ m-a.e. in Y but $\varphi \not\equiv 0$ on X. If we set $u(x,t) = \sin(\omega t)\varphi$, it is easy to see that u is a solution of (2.5) and that $u_t \not\rightharpoonup 0$ as $t \to +\infty$ weakly in H.

Remark. We deliberately chose here to a priori assume the existence of solutions on $(0, +\infty)$ rather than adding more or less classical assumptions ensuring existence of global solutions. Indeed, we concentrate here on the question of asymptotic stability. Note also that, in specific applications, existence may be provided by various means. For each of the examples below, we refer to the literature where existence results are given.

2.2. First example: A wave equation with a boundary control. In this section, we apply Theorem 2.1 to the example presented in the introduction. Note that we know in this example that global solutions of (1.2) exist at least if we suppose q continuous increasing (see, for example, [1]).

COROLLARY 2.1. *Let Ω, Γ_0, Γ^* be as in Example* 1.1. *Let $a : \Gamma_0 \to \mathbb{R}$ be such that $a(.) > 0$ on Γ_0, and let $q : \mathbb{R} \to \mathbb{R}$ be continuous and such that (H_2)(i) holds. Let (u_0, v_0) be given in $V \times L^2(\Omega)$ (where V is defined in Example* 1.1*). We suppose that there exists a solution of* (1.2) *in the following sense:*

$$\begin{cases} (u, u_t) \in \mathcal{C}([0, +\infty[; V \times L^2(\Omega)), \\ u_t \in L^2_{loc}(]0, +\infty[; V), \\ a(.)q(u_t) \in L^1_{loc}(]0, +\infty[\times\Gamma_0), \\ u_{tt} \in L^2_{loc}(]0, +\infty[; L^2(\Omega)), \end{cases} \tag{2.7}$$

and

$$\begin{cases} u_{tt} - \Delta u = 0 & \text{in } \mathcal{D}'(]0, +\infty[\times\Omega) \text{ and a.e. } (t,x) \in \mathbb{R}_+ \times \Omega, \\ \frac{\partial u}{\partial \nu} = -a(x)q(u_t) & \text{a.e. } (t,x) \in \mathbb{R}_+ \times \Gamma_0, \\ u(0) = u_0, u_t(0) = v_0 & \text{a.e. } x \in \Omega. \end{cases} \tag{2.8}$$

Then the solution satisfies

$$(u(t), u_t(t)) \rightharpoonup (0,0) \text{ weakly in } V \times L^2(\Omega) \text{ as } t \to +\infty.$$

Remark. In particular, this result is true for q continuous monotone increasing such that $q(0) = 0$ and $q(\lambda) > 0\ \forall \lambda > 0$.

Remark. This example is also treated in [10] by another method: there we use Young's measures to handle the weak limits.

2.3. Second example: A plate equation with interior control. In this section, we apply Theorem 2.1 to the following example.

EXAMPLE 2.1. *Let Ω be a connected bounded open subset of $\mathbb{R}^N$ with a smooth boundary Γ. Let $q : \mathbb{R} \to \mathbb{R}$ be continuous and such that (H_2)(i) holds, and let $a : [0,1] \to \mathbb{R}$ be such that $a(.) \in L^\infty(0,1)$, $a(.) \geq 0$, and meas$(\mathrm{supp}(a)) > 0$. We set $V = \{v \in H^2(\Omega) \mid v = 0 \text{ on } \Gamma\}$, and we consider the following problem:*

$$\begin{cases} u_{tt} + \Delta^2 u = -a(x)q(u_t) & \text{for } (t,x) \in \mathbb{R}_+ \times \Omega, \\ u = 0 & \text{on } \Gamma,\ t > 0, \\ \Delta u = 0 & \text{on } \Gamma,\ t > 0, \\ u(0) = u_0, u_t(0) = v_0 & \text{in } \Omega, \end{cases} \tag{2.9}$$

where (u_0, v_0) are given in $V \times L^2(\Omega)$.

Remark. In this case, we can prove that if we suppose q continuously differentiable such that

$$\exists m \in \mathbb{R} \mid \ \forall \lambda \in \mathbb{R},\ q'(\lambda) \geq -m,$$

then global solutions of (2.9) exist.

More generally, it is still true in the case of the abstract problem (2.5) with an interior control, i.e., when $X = Y = \Omega$ and $\mu = m = dx$ (for proofs, see [11]).

Applying Theorem 2.1 to this problem, we obtain the following result.

COROLLARY 2.2. *All global solutions of Example* 2.1 *satisfy*

$$(u(t), u_t(t)) \rightharpoonup (0,0) \text{ weakly in } V \times L^2(\Omega) \text{ as } t \to +\infty.$$

Remark. If we suppose $N = 1$, $\Omega =]0,1[$, and q globally Lipschitz continuous, then strong asymptotic stability was obtained by E. Feireisl (see [5]), i.e., $(u(t), u_t(t)) \to (0,0)$ strongly in $H^2(0,1) \times L^2(0,1)$ as $t \to +\infty$.

2.4. Third example: Rectangular Kirchhoff plates with an inner point control.

EXAMPLE 2.2. *We consider $\Omega =]0,a[\times]0,b[$ with $a > 0, b > 0$ and such that*

$$\frac{a^2}{b^2} \notin \mathbb{Q}. \tag{2.10}$$

Let $(x_0, y_0) \in \Omega$, and let $q : \mathbb{R} \to \mathbb{R}$ be continuous and such that (H_2)(i) holds. We set $V = \{v \in H^2(\Omega) \mid v = 0 \text{ on } \Gamma = \partial\Omega\}$, and we consider the following problem:

$$\begin{cases} u_{tt} + \Delta^2 u = -q(u_t(x_0,y_0))\delta(\cdot - x_0, \cdot - y_0) & \text{for } (t,x) \in \mathbb{R}_+ \times \Omega, \\ u = 0 & \text{on } \Gamma,\ t > 0, \\ \Delta u = 0 & \text{on } \Gamma,\ t > 0, \\ u(x,y,0) = u_0(x,y), u_t(x,y,0) = v_0(x,y) & \text{in } \Omega, \end{cases} \tag{2.11}$$

where (u_0, v_0) are given in $V \times L^2(\Omega)$.

This problem was studied for $q = C.Id$ in [12] and in [1], and strong asymptotic stability was obtained. Note that global solutions of (2.11) exist at least if we suppose q continuous increasing (see, for example, [1]). Applying Theorem 2.1, we obtain the following result.

COROLLARY 2.3. *If* (2.10) *is satisfied, then all global solutions of Example* 2.2 *satisfy* $(u(t), u_t(t)) \rightharpoonup (0,0)$ *weakly in* $V \times L^2(\Omega)$ *as* $t \to +\infty$ *if and only if* $\frac{x_0}{a} \notin \mathbb{Q}$ *and* $\frac{y_0}{b} \notin \mathbb{Q}$.

Remark. Theorem 2.1 can be applied to many other equations (see [10], [11] for other examples).

2.5. Remarks. It can be proved that in the case of an interior control ($X = Y = \Omega, m = \mu = dx$), the hypothesis (H_2)(i), (2.2) can be replaced by the weaker hypothesis

$$\forall \lambda > 0,\ q(\lambda) > 0$$

(for proof, see [11]). But we do not know if it is still true in the general case (and in particular in the case of a boundary control).

However, the hypothesis $\forall \lambda > 0,\ q(\lambda) > 0$, is absolutely necessary (at least in the case of a boundary control) since we have the following counterexample.

EXAMPLE 2.3. *Let* $q : \mathbb{R} \to \mathbb{R}$ *be defined by*

$$\begin{cases} q(\lambda) = 0 & \text{if } \lambda \le 0, \\ q(\lambda) = \lambda & \text{if } 0 \le \lambda \le 1, \\ q(\lambda) = 2 - \lambda & \text{if } 1 \le \lambda \le 2, \\ q(\lambda) = 0 & \text{if } 2 \le \lambda, \end{cases}$$

and let u *be the solution of*

$$\begin{cases} u_{tt} - u_{xx} = 0, & x \in]0,1[,\ t > 0, \\ u(0,t) = 0, & t > 0, \\ -u_x(1,t) = q(u_t(1,t)), & t > 0, \end{cases}$$

with

$$\begin{cases} u_0(x) = 0, & x \in]0,1[, \\ u_1(x) = 2\alpha, & x \in]0,1[. \end{cases}$$

Then, if $\alpha > 1$ *or* $\alpha < -1$, *we can prove that* $u(t) \not\to 0$ *weakly in* $H^1(0,1)$ *as* $t \to +\infty$. *To prove this, we make the calculus of the solution and we remark that the support of* u_t *is contained in the kernel of* q *(for details, see* [11]*).*

3. Proofs.

3.1. Proof of Theorem 2.1. Let (u_0, v_0) be given in $V \times H$, and let $u(t; u_0, v_0)$ be the solution of (2.5) associated with the initial data (u_0, v_0). From (2.6) and (H_1)(ii), we obtain

$$\langle u_{tt}, u_t \rangle_{V',V} + \langle \tilde{A}u + Q(u_t), u_t \rangle_{V',V} = 0 \text{ a.e. } t \in \mathbb{R}_+,$$

with

$$\langle u_{tt}, u_t \rangle_{V',V} = \frac{1}{2}\frac{d}{dt}\|u_t\|_H^2 \text{ a.e. } t \in \mathbb{R}_+,$$

and

$$\langle \tilde{A}u, u_t \rangle_{V',V} = (u, u_t)_V = \frac{1}{2}\frac{d}{dt}\|u\|_V^2 \text{ a.e. } t \in \mathbb{R}_+,$$

so that, a.e. $t \in \mathbb{R}_+$, from (2.4) we have

$$\frac{d}{dt}(\|u_t\|_H^2 + \|u\|_V^2) = -2\langle Q(u_t), u_t\rangle_{V',V} = -2\int_Y aq(u_t)u_t dm. \tag{3.1}$$

Hence, we obtain the following "energy" equality:

$$\|u(t)\|_V^2 + \|u_t(t)\|_H^2 + 2\int_0^t \int_Y aq(u_t)u_t dmds = \|u_0\|_V^2 + \|v_0\|_H^2 \text{ a.e. } t \in \mathbb{R}_+. \tag{3.2}$$

If we set $v = u_t$ and $U = \left(\begin{smallmatrix} u \\ v \end{smallmatrix}\right)$, then the problem (2.5) may be written in the first order form

$$\begin{cases} \frac{dU}{dt} = \mathcal{A}U + F(U), & t \in \mathbb{R}_+, \\ U(t) \in \mathcal{H}, & t \in \mathbb{R}_+, \\ U(0) = U_0, \end{cases} \tag{3.3}$$

where

$$\mathcal{A} = \begin{pmatrix} 0 & 1 \\ -\tilde{A} & 0 \end{pmatrix},\ F(U) = \begin{pmatrix} 0 \\ -Q(v) \end{pmatrix},\ U_0 = \begin{pmatrix} u_0 \\ v_0 \end{pmatrix},$$

and where $\mathcal{H} = V \times H$ is equipped with the scalar product

$$(U, \tilde{U})_{\mathcal{H}} = (u, \tilde{u})_V + (v, \tilde{v})_H \text{ for } U = \begin{pmatrix} u \\ v \end{pmatrix},\ \tilde{U} = \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \in \mathcal{H}.$$

We also note that

$$(U, \tilde{U})_{\mathcal{H}' \times \mathcal{H}} = \langle u, \tilde{u}\rangle_{V',V} + (v, \tilde{v})_H$$
$$\text{for } U = \begin{pmatrix} u \\ v \end{pmatrix} \in \mathcal{H}' = V' \times H \text{ and } \tilde{U} = \begin{pmatrix} \tilde{u} \\ \tilde{v} \end{pmatrix} \in \mathcal{H} = V \times H.$$

With these notations, (3.2) becomes

$$\|U(t; U_0)\|_{\mathcal{H}}^2 + 2\int_0^t \int_Y aq(u_t)u_t dmds = \|U_0\|_{\mathcal{H}}^2 \text{ a.e. } t \in \mathbb{R}_+. \tag{3.4}$$

By (2.1) and (3.4), we have

$$\|U(t; U_0)\|_{\mathcal{H}} \le \|U_0\|_{\mathcal{H}} \text{ a.e. } t \in \mathbb{R}_+. \tag{3.5}$$

Consequently, the weak ω-limit set $\omega_W(U_0)$ is nonempty. Let $\begin{pmatrix} \varphi \\ \psi \end{pmatrix}$ be in $\omega_W(U_0)$. By definition, there exists $t_n \to +\infty$ as $n \to +\infty$ such that

$$U(t_n; U_0) \rightharpoonup \begin{pmatrix} \varphi \\ \psi \end{pmatrix} \text{ in } \mathcal{H} \text{ as } n \to +\infty. \tag{3.6}$$

For this sequence t_n, we consider the translates defined by

$$U_n(t) = U(t + t_n; U_0). \tag{3.7}$$

For any fixed $T > 0$, we deduce from (3.5) that $(U_n)_n$ is a bounded sequence in $L^\infty(0, T; \mathcal{H})$. Hence it possesses a subsequence also denoted by $(U_n)_n$, and there exists $\bar{U} = \left(\begin{smallmatrix} \bar{u} \\ \bar{v} \end{smallmatrix}\right) \in L^\infty(0, T; \mathcal{H})$ such that

$$U_n = \begin{pmatrix} u_n \\ v_n \end{pmatrix} \rightharpoonup \bar{U} = \begin{pmatrix} \bar{u} \\ \bar{v} \end{pmatrix} \text{ weakly in } L^2(0, T; \mathcal{H}) \text{ as } n \to +\infty. \tag{3.8}$$

In particular, we can prove that for any $W \in \mathcal{H}$, the function $t \mapsto (U_n(t), W)_{\mathcal{H}}$ converges in $\mathcal{D}'(]0,T[)$ to the function $t \mapsto (\bar{U}(t), W)_{\mathcal{H}}$.

Our purpose is to prove that $\bar{U} \equiv 0$ on $X \times \mathbb{R}_+$. In particular, it will give $\bar{U}(0) = \binom{\varphi}{\psi} = \binom{0}{0}$ and so $\omega_W(U_0) = \{(0,0)\}$, which implies that $U(t;U_0) \rightharpoonup (0,0)$ weakly in $V \times H$. In the first part of the proof, we will write the equation satisfied by the translates U_n. Passing to the limit as $n \to +\infty$ in this equation, we will obtain an equation satisfied by $\bar{U}$. In the second part, we will use the representation of the solutions of this equation and the result of "uniqueness" (U) to prove that $\bar{U} \equiv 0$ on $X \times \mathbb{R}_+$.

Part 1. Since U_n is a solution of (3.3) with the initial data $U(t_n)$, we have for all $n \in \mathbb{N}$, $t \in [0,T]$, and $W = \binom{w_1}{w_2} \in V \times \mathcal{E}$,

$$(U_n(t), W)_{\mathcal{H}} - (U(t_n), W)_{\mathcal{H}} = \int_0^t (\mathcal{A}^* W, U_n(s))_{\mathcal{H}' \times \mathcal{H}} ds + \int_0^t \int_Y F(U_n(s)) \cdot W dm(y) ds. \tag{3.9}$$

By (3.8), the first term $(U_n(t), W)_{\mathcal{H}}$ converges in $\mathcal{D}'(]0,T[)$ to $(\bar{U}(t), W)_{\mathcal{H}}$ as $n \to +\infty$. By (3.6),

$$(U(t_n), W)_{\mathcal{H}} \to \left(\binom{\varphi}{\psi}, W \right)_{\mathcal{H}} \text{ as } n \to +\infty,$$

and by (3.8), we have for all $t \in [0,T]$,

$$\int_0^t (\mathcal{A}^* W, U_n(s))_{\mathcal{H}' \times \mathcal{H}} ds \to \int_0^t (\mathcal{A}^* W, \bar{U}(s))_{\mathcal{H}' \times \mathcal{H}} ds \text{ as } n \to +\infty.$$

Moreover, the Lebesgue dominated convergence theorem implies that these two last convergences hold in $\mathcal{D}'(]0,T[)$.

We will now prove that

$$a(y) q(v_n) \to 0 \text{ as } n \to +\infty \text{ strongly in } L^1(Y \times [0,T]), \tag{3.10}$$

which will imply that the last term of (3.9) converges to 0 in $\mathcal{D}'(]0,T[)$. Indeed, we will obtain for all $t \in [0,T]$,

$$\int_0^t \int_Y F(U_n(s)) \cdot W dm(y) ds = \begin{pmatrix} 0 \\ -\int_0^t \int_Y a(y) q(v_n) w_2 dm(y) ds \end{pmatrix} \to \binom{0}{0} \text{ as } n \to +\infty,$$

and the Lebesgue dominated convergence theorem will give the result since w_2 is bounded in Y.

To prove (3.10), we fix $\varepsilon > 0$. By continuity of $q(\lambda)$ at $\lambda = 0$, there exists $\eta(\varepsilon)$ such that

$$\iint_{\substack{Y \times [0,t] \\ |v_n(y,s)| \le \eta(\varepsilon)}} a(y) |q(v_n)| dm(y) ds \le \varepsilon. \tag{3.11}$$

So, we have

$$\begin{aligned} &\int_0^t \int_Y a(y) |q(v_n)| dm(y) ds \\ &\le \iint_{\substack{Y \times [0,t] \\ |v_n(y,s)| \ge \eta(\varepsilon)}} a(y) |q(v_n)| dm(y) ds + \iint_{\substack{Y \times [0,t] \\ |v_n(y,s)| \le \eta(\varepsilon)}} a(y) |q(v_n)| dm(y) ds \\ &\le \frac{1}{\eta(\varepsilon)} \int_0^t \int_Y a(y) q(v_n) v_n dm(y) ds + \varepsilon. \end{aligned} \tag{3.12}$$

From (3.4), we have for all $t \in [0,T]$ and $n \in \mathbb{N}$,

$$\|U_n(t)\|_{\mathcal{H}}^2 - \|U(t_n)\|_{\mathcal{H}}^2 = -2\int_0^t \int_Y a(y)q(v_n)v_n dm(y)ds. \tag{3.13}$$

Since the function $t \mapsto \|U(t;U_0)\|_{\mathcal{H}}$ is nonincreasing and bounded from below, its limit as $t \to +\infty$ exists. But since $U_n(t) = U(t+t_n)$, it follows that

$$\lim_{n\to+\infty} \|U_n(t)\|_{\mathcal{H}} = \lim_{n\to+\infty} \|U(t_n;U_0)\|_{\mathcal{H}},$$

and hence by (3.13),

$$\lim_{n\to+\infty} \int_0^t \int_Y a(y)q(v_n)v_n dm(y)ds = 0 \quad \forall t \in [0,T]. \tag{3.14}$$

We deduce from (3.12) and (3.14) that there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$,

$$\int_0^t \int_Y a(y)|q(v_n)|dm(y)ds \leq 2\varepsilon.$$

So (3.10) is proved.

We can now write the limit in $\mathcal{D}'(]0,T[)$ of relation (3.9) as $\forall t \in [0,T]$, $\forall W = \left(\begin{smallmatrix} w_1 \\ w_2 \end{smallmatrix}\right) \in V \times \mathcal{E}$,

$$(\bar{U}(t),W)_{\mathcal{H}} - \left(\left(\begin{smallmatrix} \varphi \\ \psi \end{smallmatrix}\right), W\right)_{\mathcal{H}} = \int_0^t (\mathcal{A}^*W, \bar{U}(s))_{\mathcal{H}'\times\mathcal{H}} ds. \tag{3.15}$$

And we obtain that $\bar{U}$ is a solution of

$$\begin{cases} \frac{d}{dt}(\bar{U}(t),W)_{\mathcal{H}} = (\mathcal{A}^*W, \bar{U}(t))_{\mathcal{H}\times\mathcal{H}} \quad \forall W \in D(\mathcal{A}^*) \forall t \in [0,T], \\ \bar{U}(0) = \left(\begin{smallmatrix} \varphi \\ \psi \end{smallmatrix}\right). \end{cases} \tag{3.16}$$

Indeed, (3.15) implies that this equation is verified for all $W \in V \times \mathcal{E}$, and since $\mathcal{E}$ is dense in $D(A)$, it is still true for all $W \in V \times D(A) = D(\mathcal{A}^*)$. Moreover, since T is arbitrary and $\bar{U} \in L^\infty(0,\infty;\mathcal{H})$, it is always true for $t \in [0,+\infty[$. Finally, $\bar{U}$ is a solution of the equation

$$\begin{cases} \frac{d}{dt}\bar{U}(t) = \mathcal{A}\bar{U}(t), \qquad t \in \mathbb{R}_+, \\ \bar{U}(0) = \left(\begin{smallmatrix} \varphi \\ \psi \end{smallmatrix}\right), \end{cases} \tag{3.17}$$

and consequently, $\bar{u}$ is a solution of

$$\begin{cases} \bar{u}_{tt} + \tilde{A}\bar{u} = 0, \qquad t \in \mathbb{R}_+, \\ \bar{u}(0) = \varphi, \ \bar{u}_t(0) = \psi. \end{cases} \tag{3.18}$$

Part 2.

Step 1. Since $(\bar{u}, \bar{u}_t)$ is a solution of (3.18), and since A satisfies (H_1)(ii), we can write

$$\bar{u}(t) = \sum_{p\in\mathbb{N}} \left(-\cos(\omega_p t)\frac{\psi_p}{\omega_p} + \sin(\omega_p t)\frac{\varphi_p}{\omega_p}\right), \tag{3.19}$$

and

$$\bar{u}_t(t) = \sum_{p\in\mathbb{N}} \left(\cos(\omega_p t)\varphi_p + \sin(\omega_p t)\psi_p\right), \tag{3.20}$$

where φ_p, ψ_p are eigenfunctions of A, ω_p eigenvalues of A and where (3.19) (resp., (3.20)) converges in V (resp., in H) uniformly in t (for more details, see [1] or [11]).

Step 2. Fix $p \in \mathbb{N}$ and $\eta = \pm 1$. For $T > 0$, we set

$$\forall n \in \mathbb{N},\ w_{n,T} = \frac{1}{T}\int_0^T \frac{1+\eta\cos(\omega_p t)}{2} v_n(x,t)dt, \tag{3.21}$$

and

$$\bar{w}_T = \frac{1}{T}\int_0^T \frac{1+\eta\cos(\omega_p t)}{2} \bar{u}_t(x,t)dt. \tag{3.22}$$

The following convergence results can be proved (the proof appears later in this paper).

LEMMA 3.1.

$$\forall T > 1,\ w_{n,T}(y) \rightharpoonup \bar{w}_T(y) \text{ weakly in } V \text{ as } n \to +\infty, \tag{3.23}$$

$$\bar{w}_T(y) \rightharpoonup \frac{1}{4}\eta\varphi_p(y) \text{ weakly in } V \text{ as } T \to +\infty. \tag{3.24}$$

Step 3. Since q is such that (H_2)(i) holds, we can construct $\hat{q} : \mathbb{R} \to \mathbb{R}$ continuous increasing such that

$$\begin{cases} \forall \lambda \in \mathbb{R}, & 0 \le \hat{q}(\lambda)\lambda \le q(\lambda)\lambda, \\ \forall \lambda > 0, & \hat{q}(\lambda) > 0 \end{cases}$$

(indeed, we can choose $\hat{q}(\lambda) = 0$ for $\lambda \le 0$ and $\hat{q}(\lambda) = \inf\{q(s) \mid s \ge \lambda\}$ for $\lambda > 0$).

Then we denote by $\hat{\jmath} : \mathbb{R} \to [0, +\infty[$ the primitive of $\hat{q}$ such that $\hat{\jmath}(0) = 0$, and we set $\hat{\Psi}(v) = \int_Y a(y)\hat{\jmath}(v(y))dm(y)$ for $v \in V$.

$\hat{\Psi} : V \to]-\infty, +\infty]$ is proper, convex (since $\hat{\jmath}$ is convex), lower semicontinuous (l.s.c.). Indeed, since $\hat{\Psi}$ is convex, we just have to prove that

$$(v_n \to v \text{ strongly in } V \text{ as } n \to \infty) \implies (\hat{\Psi}(v) \le \liminf_{n\to+\infty} \hat{\Psi}(v_n)). \tag{3.25}$$

By (H_1)(iii), $v_n \to v$ strongly in V as $n \to +\infty$ implies that $v_n \to v$ strongly in $L^1(Y, m)$ as $n \to +\infty$ and so $v_n \to v$ m-a.e. in Y as $n \to +\infty$ for a subsequence still denoted by $(v_n)_n$. Then (3.25) is verified by Fatou's lemma.

Since $\forall \lambda \in \mathbb{R}, \hat{\jmath}(\lambda) \le \lambda\hat{q}(\lambda)$, we can prove that

$$\forall n \in \mathbb{N},\ 0 \le \hat{\Psi}(v_n) \le \int_Y a\hat{q}(v_n)v_n dm \le \int_Y aq(v_n)v_n dm,$$

which gives, for all $T > 0$,

$$\forall n \in \mathbb{N},\ \int_0^T \hat{\Psi}(v_n)dt \le \int_0^T \int_Y aq(v_n)v_n dmdt.$$

So from (3.14), we deduce that

$$\forall T>0,\ \int_0^T \hat{\Psi}(v_n)dt \to 0 \text{ as } n\to+\infty. \tag{3.26}$$

Moreover, since $\hat{\Psi}$ is convex with $\hat{\Psi}(0)=0$ and since $0\leq \frac{1+\eta\cos(\omega_p t)}{2}\leq 1$ for all $t\in[0,T]$, we have

$$\forall n\in\mathbb{N},\ 0\leq \hat{\Psi}\left(\frac{1}{T}\int_0^T \frac{1+\eta\cos(\omega_p t)}{2}v_n dt\right)\leq \frac{1}{T}\int_0^T \frac{1+\eta\cos(\omega_p t)}{2}\hat{\Psi}(v_n)dt$$
$$\leq \frac{1}{T}\int_0^T \hat{\Psi}(v_n)dt.$$

Therefore, by (3.26), all these integrals tend to 0 as $n\to+\infty$. From Lemma 3.1, and since $\hat{\Psi}$ is l.s.c. on V, we have for all $T>0$,

$$\hat{\Psi}(\bar{w}_T)\leq \liminf_{n\to+\infty}\hat{\Psi}(w_{n,T})=0,$$

and similarly

$$\hat{\Psi}\left(\frac{\eta}{4}\varphi_p\right)\leq \liminf_{T\to+\infty}\hat{\Psi}(\bar{w}_T)=0.$$

So, we obtain

$$\hat{\Psi}\left(\frac{\eta}{4}\varphi_p\right)=\int_Y a(y)\hat{\jmath}\left(\frac{\eta}{4}\varphi_p(y)\right)dm(y)=0,$$

which implies

$$a(y)\hat{\jmath}\left(\frac{\pm 1}{4}\varphi_p(y)\right)=0 \text{ m-a.e. on } Y. \tag{3.27}$$

Since $\hat{\jmath}'(\lambda)=\hat{q}(\lambda)>0\ \forall\lambda>0$, we know that $\hat{\jmath}$ is strictly increasing on $\mathbb{R}_+$. But $\hat{\jmath}(0)=0$, so $\hat{\jmath}(\lambda)>0\ \forall\lambda>0$. In particular, $\hat{\jmath}(\lambda)=\hat{\jmath}(-\lambda)=0$ implies $\lambda=0$. So, (3.27) implies that

$$a(y)\varphi_p(y)=0 \text{ m-a.e. on } Y.$$

And from the "uniqueness" hypothesis (U), we obtain

$$\varphi_p(x)=0 \text{ a.e. } x\in X\ \forall p\in\mathbb{N}. \tag{3.28}$$

Similarly, multiplying v_n by $\frac{1+\eta\sin(\omega_p t)}{2}$ instead of $\frac{1+\eta\cos(\omega_p t)}{2}$, we get

$$\psi_p(x)=0 \text{ a.e. } x\in X\ \forall p\in\mathbb{N}. \tag{3.29}$$

Finally, from (3.19) and (3.20), we get $(\bar{u},\bar{u}_t)\equiv(0,0)$ on $X\times\mathbb{R}_+$, which ends the proof of Theorem 2.1. □

Proof of Lemma 3.1. First, we will prove that

$$\exists C_p>0 \text{ such that } \forall n\in\mathbb{N},\ \forall T>1,\ \|w_{n,T}\|_V\leq C_p. \tag{3.30}$$

Indeed, for all $n \in \mathbb{N}$ and $T > 0$,

$$w_{n,T}(x) = \frac{1}{T}\int_0^T \frac{1+\eta\cos(\omega_p t)}{2} v_n(x,t)\mathrm{dt}$$
$$= \frac{1}{T}\left[\frac{1+\eta\cos(\omega_p t)}{2} u_n(x,t)\right]_0^T + \frac{1}{T}\int_0^T \frac{\eta\omega_p \sin(\omega_p t)}{2} u_n(x,t)\mathrm{dt}.$$

Hence,

$$\|w_{n,T}\|_V \leq \frac{1}{T}\left(\frac{1+\eta\cos(\omega_p t)}{2}\|u_n(T)\|_V + \frac{1+\eta}{2}\|u_n(0)\|_V\right) + \frac{1}{T}\int_0^T \frac{1}{2}\omega_p \|u_n(t)\|_V \mathrm{dt}.$$

From (3.5), we know that there exists $C > 0$ such that

$$\|u_n(t)\|_V \leq C \ \forall n \in \mathbb{N},\ \forall t \in \mathbb{R}_+.$$

Consequently,

$$\|w_{n,T}\|_V \leq \frac{2C}{T} + \frac{1}{2}\omega_p C \leq 2C + \frac{1}{2}\omega_p C = C_p \ \text{ for } T > 1,$$

which proves (3.30).

From (3.30), we deduce that for $T > 1$, $w_{n,T}$ converges weakly in V as $n \to +\infty$ (for a subsequence still denoted by $w_{n,T}$). But $v_n \rightharpoonup \bar{v} = \bar{u}_t$ weakly in $L^2(0,T;H)$ as $n \to +\infty$, hence for all $T > 1$, we have

$$w_{n,T} \rightharpoonup \bar{w}_T \text{ weakly in } H \text{ as } n \to +\infty.$$

So by uniqueness of the limit for the weak convergence in H, we obtain

$$w_{n,T} \rightharpoonup \bar{w}_T \text{ weakly in } V \text{ as } n \to +\infty, \tag{3.31}$$

which proves the first part of Lemma 3.1.

Using (3.31) and (3.30), we obtain

$$\|\bar{w}_T\| \leq C_p \text{ for all } T > 1. \tag{3.32}$$

We deduce that $\bar{w}_T$ converges weakly in V as $T \to +\infty$ (for a subsequence still denoted $\bar{w}_T$). But it can be proved that, for $\eta = \pm 1$ and for all $p \in \mathbb{N}$,

$$\frac{1}{T}\int_0^T \frac{1+\eta\cos(\omega_p t)}{2}\bar{u}_t(t)\mathrm{dt} \longrightarrow \frac{1}{4}\eta\varphi_p \text{ strongly in } H \text{ as } T \to +\infty, \tag{3.33}$$

$$\text{i.e., } \bar{w}_T \longrightarrow \frac{1}{4}\eta\varphi_p \text{ strongly in } H \text{ as } T \to +\infty.$$

This convergence is a consequence of the fact that (3.20) converges in H uniformly in t. It consists as in [1] in multiplying (3.20) by $\frac{1+\eta\cos(\omega_p t)}{2}$ and computing the limit. In [1], (3.20) converges in V uniformly in t and the convergence (3.33) is proved in V. Here the convergence is only true in H, but the proof is the same (for details, see [11]). So by uniqueness of the limit, we deduce that

$$\bar{w}_T \rightharpoonup \frac{1}{4}\eta\varphi_p \text{ weakly in } V \text{ as } T \to +\infty,$$

which proves the second part of Lemma 3.1. □

3.2. Remark concerning the definition of Q. In the following examples, for all $v \in D(Q)$, $Q(v)$ will be defined on $\mathcal{E}$ by

$$\forall \varphi \in \mathcal{E},\ \langle Q(v), \varphi \rangle_{V',V} = \int_Y a(y) q(v(y)) \varphi(y) dm(y) \tag{3.34}$$

(see (H_2)(ii)).

$Q(v)$ is linear continuous from $\mathcal{E}$ into $\mathbb{R}$ and it can be extended to a linear continuous application from V into $\mathbb{R}$ still denoted $Q(v)$. So for all $v \in D(Q)$, $Q(v)$ is defined and $Q(v) \in V'$. But in order to verify hypothesis (H_2)(ii), we still have to verify that equation (3.34) is still true for $\varphi = v$, i.e., we have to prove that

$$\forall v \in D(Q),\ \langle Q(v), v \rangle_{V',V} = \int_Y a(y) q(v(y)) v(y) dm(y). \tag{3.35}$$

In this paper, we will omit this technical point in the proofs of Corollaries 2.1, 2.2, and 2.3. It is generally based on "smooth truncation" results in Sobolev spaces. For details, see [11].

Note that, in the proof of Theorem 2.1, relation (3.35) is only used in (3.1) in order to obtain the "energy equality" (3.2). In most examples, if we add assumptions ensuring existence of global solutions, the "energy equality" (3.2) will be obtained naturally, and consequently, it will not be necessary to verify (3.35).

3.3. Proof of Corollary 2.1. We set the following:

- $(X, \mu) = (\overline{\Omega}, dx)$ where dx is the Lebesgue's measure on $\mathbb{R}^N$.
  $H = L^2(\Omega, dx)$.
- $V = \{v \in H^1(\Omega) \mid v_{|\Gamma^*} = 0\}$.
  $D(A) = \{u \in V \mid \Delta u \in L^2(\Omega) \text{ and } \frac{\partial u}{\partial \nu} = 0 \text{ on } \Gamma_0\}$ (note that $\frac{\partial u}{\partial \nu}_{|\Gamma}$ is defined and $\frac{\partial u}{\partial \nu}_{|\Gamma} \in H^{-\frac{1}{2}}(\Gamma)$ because $u \in H^1(\Omega)$ with $\Delta u \in L^2(\Omega)$, see [8] or [4]).
  For $u \in D(A)$, $Au = -\Delta u$, and for $(u, v) \in V \times V$, $\langle \tilde{A}u, v \rangle_{V',V} = \int_\Omega \nabla u \nabla v dx$.
- $(Y, m) = (\Gamma, d\sigma)$ where $d\sigma$ is the superficial measure on Γ.
  The trace mapping $\tau : V \to L^1(\Gamma, d\sigma)$ is continuous (it is actually continuous into $H^{\frac{1}{2}}(\Gamma, d\sigma)$) .
- $a(.)$ is defined on Γ_0 and we extend it in a function still denoted by $a(.)$ and defined on Γ by $a(x)$ for $x \in \Gamma_0$ and 0 for $x \in \Gamma^*$.

Then we can apply Theorem 2.1, and weak asymptotic stability is obtained if the following "uniqueness" result is true:

$$\left(-\Delta \varphi = \omega^2 \varphi \text{ in } \Omega,\ \varphi = 0 \text{ on } \Gamma^*,\ \frac{\partial \varphi}{\partial \nu} = 0 \text{ on } \Gamma_0 \textbf{ and } \varphi = 0 \text{ on } \Gamma_0 \right) \Longrightarrow \varphi \equiv 0 \text{ in } \Omega.$$

This result is true because Γ_0 is not too "thin" (for example Γ_0 contains $B(x_0, \varepsilon) \cap \Gamma$ where $x_0 \in \Gamma$, see [1]). □

3.4. Proof of Corollary 2.2. We set the following:

- $(X, \mu) = (Y, m) = (\Omega, dx)$ where dx is the Lebesgue's measure on $\mathbb{R}^N$.
  $H = L^2(\Omega, dx)$.
- $V = \{v \in H^2(\Omega) \mid v = 0 \text{ on } \Gamma\}$.

$D(A) = \{u \in V \mid \Delta^2 u \in L^2(\Omega)$ and $\Delta u_{|\Gamma} = 0$ on $\Gamma\}$ (note that $\Delta u_{|\Gamma}$ is defined and $\Delta u_{|\Gamma} \in H^{-\frac{1}{2}}(\Gamma)$ because $\Delta u \in L^2(\Omega)$ with $\Delta^2 u \in L^2(\Omega)$, see [8]).

For $u \in D(A)$, $Au = \Delta^2 u$ and for $(u, v) \in V \times V$, $\langle \tilde{A}u, v\rangle_{V',V} = \int_\Omega \Delta u \Delta v dx$.

Then we can apply Theorem 2.1 and weak asymptotic stability is obtained because the following "uniqueness" result is true:

$$\big(\varphi \in V,\ \Delta^2\varphi = \omega^2\varphi \text{ in } \Omega, \quad \textbf{and } \varphi(x) = 0 \text{ a.e. } x \in \mathrm{supp}(a)\big) \Longrightarrow \varphi \equiv 0 \text{ in } \Omega.$$

Indeed it implies that φ is analytic on Ω (see [4]), and consequently, $\varphi \equiv 0$ in Ω since Ω is connected and $\mathrm{meas}(\mathrm{supp}(a)) > 0$. □

3.5. Proof of Corollary 2.3. We set the following:

- $(X, \mu) = (\Omega, dxdy) = (]0, a[\times]0, b[, dxdy)$ where $dxdy$ is the Lebesgue's measure in $\mathbb{R}^2$.
  $H = L^2(]0, a[\times]0, b[, dxdy)$.
- $V = \{v \in H^2(\Omega) \mid v = 0 \text{ on } \Gamma = \partial\Omega\}$.
  $D(A) = \{u \in V \mid \Delta^2 u \in H$ and $\Delta u = 0$ on $\Gamma\}$.
  For $u \in D(A)$, $Au = \Delta^2 u$ and for $(u, v) \in V \times V$, $\langle \tilde{A}u, v\rangle_{V',V} = \int_\Omega \Delta u \Delta v dxdy$.
- $Y = \{(x_0, y_0)\}$ and m is the Dirac masse $\delta(\cdot - x_0, \cdot - y_0)$.
- We define $a : Y \to \mathbb{R}$ by $a(x_0, y_0) = 1$.

Then we can verify that (H_1) and (H_2) are satisfied. Indeed, A is self-adjoint, coercive on H, and the resolvent of A is compact (see, for example, [12]). Moreover, $V \subset H^2(\Omega) \hookrightarrow \mathcal{C}^0(\overline{\Omega})$ (because $\Omega \subset \mathbb{R}^2$) so that $\forall v \in V$, $v(x_0, y_0)$ exists and $v \mapsto v(x_0, y_0)$ is linear continuous from $V \to \mathbb{R}$. Finally, Theorem 2.1 can be applied and we obtain weak asymptotic stability of all global solutions if and only if the following "uniqueness" result is true:

$$\big(\varphi \in V,\ A\varphi = \omega^2\varphi \text{ on } \Omega \textbf{ and } \varphi(x_0, y_0) = 0\big) \Longrightarrow \varphi \equiv 0 \text{ on } \Omega. \tag{3.36}$$

As in [1], we can prove that (3.36) is verified as follows: let $\varphi \in V$ be such that $A\varphi = \omega^2\varphi$ on Ω. Then, we have

$$\varphi(x, y) = \sum_{r,s} \alpha_{r,s} \sin\frac{r\pi x}{a} \sin\frac{s\pi y}{b},$$

where $r, s \in \mathbb{N}$ are such that $\omega^4 = (\frac{r^2}{a^2} + \frac{s^2}{b^2})^2\pi^4$. But the eigenvalues are all simple because $\frac{a^2}{b^2} \notin \mathbb{Q}$. Indeed, if ω is not a simple eigenvalue, then there exists $(m, n) \in \mathbb{N}^2$ and $(m', n') \in \mathbb{N}^2$ such that $(m', n') \neq (m, n)$ and $\omega^2 = (\frac{m^2}{a^2} + \frac{n^2}{b^2})\pi^2 = (\frac{m'^2}{a^2} + \frac{n'^2}{b^2})\pi^2$. So, we obtain $\frac{m^2 - m'^2}{n'^2 - n^2} = \frac{a^2}{b^2} \notin \mathbb{Q}$, which is absurd. So the hypothesis (2.10) implies that

$$\varphi(x, y) = \alpha_{m,n} \sin\frac{m\pi x}{a} \sin\frac{n\pi y}{b},$$

with $\omega^2 = (\frac{r^2}{a^2} + \frac{s^2}{b^2})\pi^2$. Clearly, $\sin\frac{m\pi x_0}{a} \sin\frac{n\pi y_0}{b} \neq 0$ if and only if $\frac{x_0}{a} \notin \mathbb{Q}$ and $\frac{y_0}{b} \notin \mathbb{Q}$, so (3.36) is true if and only if $\frac{x_0}{a} \notin \mathbb{Q}$ and $\frac{y_0}{b} \notin \mathbb{Q}$, which ends the proof. □

Acknowledgments. The author is grateful to Professor M. Pierre for first suggesting this problem and for a number of instructive discussions. The author would also like to thank Professors V. Komornik and B. Rao for their comments and suggestions on this work.

REFERENCES

[1] F. Conrad and M. Pierre, *Stabilization of second order evolution equations by unbounded nonlinear feedbacks*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 11 (1994), pp. 485–515.

[2] F. Conrad and M. Pierre, *Stabilization of Euler-Bernoulli Beam by Nonlinear Boundary Feedback*, INRIA Report RR 1235, INRIA Lorraine, Nancy, 1990.

[3] C. M. Dafermos, *Asymptotic behavior of solutions of evolutions equations*, in Nonlinear Evolution Equations, M. G. Crandall, ed., Academic Press, New York, 1978, pp. 103–123.

[4] R. Dautray and J. L. Lions, *Analyse mathématique et calcul numérique pour les sciences et techniques*, Masson, Paris, 1984.

[5] E. Feireisl, *Strong asymptotic stability for a beam equation with weak damping*, Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 365–371.

[6] A. Haraux, *Stabilization of trajectories for some weakly damped hyperbolic equations*, J. Differential Equations, 59 (1985), pp. 145–154.

[7] A. Haraux, *Comportement à l'infini pour certains systèmes dissipatifs non linéaires*, Proc. Roy. Soc. Edinburgh Sect. A, 84 (1979), pp. 213–234.

[8] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Dunod et Gauthier-Villars, Paris, 1968.

[9] M. Slemrod, *Weak asymptotic decay via a relaxed invariance principle for a wave equation with nonlinear, nonmonotone damping*, Proc. Roy. Soc. Edinburgh Sect. A, 113 (1989), pp. 87–97.

[10] J. Vancostenoble, *Stabilisation faible de l'équation des ondes par un controle non linéaire, non monotone*, Les prépublications de l'Institut Elie Cartan, No. 3, Inst. Elie Cartan University Nancy I, 1997.

[11] J. Vancostenoble, *Stabilité asymptotique faible d'équations d'évolution du second ordre par des contrôles non linéaires et non monotones*, Ph.D. thesis, University Rennes I, December 1998.

[12] Y. You, *Controllability and stabilization of vibrating simply supported plate with pointwise control*, Adv. in Appl. Math., 10 (1989), pp. 324–343.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 155–189

APPROXIMATION BY RIDGE FUNCTIONS AND NEURAL NETWORKS*

PENCHO P. PETRUSHEV†

Abstract. We investigate the efficiency of approximation by linear combinations of ridge functions in the metric of $L_2(\mathbf{B}^d)$ with $\mathbf{B}^d$ the unit ball in $\mathbf{R}^d$. If X_n is an n-dimensional linear space of univariate functions in $L_2(I)$, $I = [-1,1]$, and Ω is a subset of the unit sphere $\mathbf{S}^{d-1}$ in $\mathbf{R}^d$ of cardinality m, then the space $Y_n := \text{span}\{r(\mathbf{x}\cdot\xi) : r \in X_n, \omega \in \Omega\}$ is a linear space of ridge functions of dimension $\le mn$. We show that if X_n provides order of approximation $O(n^{-r})$ for univariate functions with r derivatives in $L_2(I)$, and Ω are properly chosen sets of cardinality $O(n^{d-1})$, then Y_n will provide approximation of order $O(n^{-r-d/2+1/2})$ for every function $f \in L_2(\mathbf{B}^d)$ with smoothness of order $r+d/2-1/2$ in $L_2(\mathbf{B}^d)$. Thus, the theorems we obtain show that this form of ridge approximation has the same efficiency of approximation as other more traditional methods of multivariate approximation such as polynomials, splines, or wavelets. The theorems we obtain can be applied to show that a feed-forward neural network with one hidden layer of computational nodes given by certain sigmoidal function σ will also have this approximation efficiency. Minimal requirements are made of the sigmoidal functions and in particular our results hold for the unit-impulse function $\sigma = \chi_{[0,\infty)}$.

Key words. approximation error, ridge functions, neural networks

AMS subject classifications. 41A15, 41A25, 41A29

PII. S0036141097322959

1. Introduction.

A ridge function is a multivariate function of the form $r(\mathbf{x}\cdot\omega)$, where r is a univariate function, ω is a fixed vector in $\mathbf{R}^d$, the variable $\mathbf{x} \in \mathbf{R}^d$, and $\mathbf{x}\cdot\omega$ is the inner product of $\mathbf{x}$ and ω. These functions appear naturally in harmonic analysis, special function theory, and in several applications such as tomography and neural networks. In most applications, we are interested in representing or approximating a general function f on a domain $\mathbf{\Omega} \subset \mathbf{R}^d$ by linear combinations of ridge functions. It is surprising therefore that the most fundamental questions concerning the efficiency of approximation by ridge functions are unanswered.

In this paper, we shall consider approximating functions in $L_2(\mathbf{B}^d)$, $\mathbf{B}^d$ the unit ball in $\mathbf{R}^d$, $d \ge 2$, by linear combinations of ridge functions. Using extension theorems, the set $\mathbf{B}^d$ can be replaced by more general sets $\mathbf{\Omega} \subset \mathbf{R}^d$.

Let X_n be a linear space of univariate functions in $L_2(I)$, $I := [-1,1]$ and let $\Omega_n \subset \mathbf{S}^{d-1}$ be a finite subset of the unit sphere $\mathbf{S}^{d-1}$ in $\mathbf{R}^d$. Then

$$Y_n := \text{span}\{r(\mathbf{x}\cdot\omega) : r \in X_n,\ \omega \in \Omega_n\} \tag{1.1}$$

is a space of multivariate ridge functions of dimension $\le n\#\Omega_n$, where $\#\Omega_n$ is the cardinality of Ω_n. We shall relate the approximation efficiency of Y_n to that of X_n and the distribution of the vectors of Ω_n in $\mathbf{S}^{d-1}$.

Let $W^s(L_2(I))$ denote the univariate Sobolev spaces. We say that a sequence of spaces $X_n, n = 1,2,\dots,$ $\dim(X_n) = n$, provides *approximation of order* s if

$$E(g,X_n)_{L_2(I)} \le c(s)n^{-s}\|g\|_{W^s(L_2(I))}, \quad g \in W^s(L_2(I)), \tag{1.2}$$

* Received by the editors June 13, 1997; accepted for publication (in revised form) February 5, 1998; published electronically October 20, 1998. This research was supported by ONR Research Contracts N00014-96-1-1003 and DAAG55-98-1-0002.

http://www.siam.org/journals/sima/30-1/32295.html

†Department of Mathematics, University of South Carolina, Columbia, SC 29208 (pencho@math.sc.edu).

where

$$E(g, X_n)_{L_2(I)} := \inf_{r \in X_n} \|g - r\|_{L_2(I)}$$

is the error in approximating the univariate function g in the $L_2(I)$ norm by the elements of X_n. We denote similarly the multivariate Sobolev space $W^s(L_2(\mathbf{B}^d))$ on $\mathbf{B}^d$ and the approximation error

$$E(f, Y_n)_{L_2(\mathbf{B}^d)} := \inf_{R \in Y_n} \|f - R\|_{L_2(\mathbf{B}^d)}$$

for any $f \in L_2(\mathbf{B}^d)$. Our main result, given in section 8, shows that for any sequence of spaces $X_n, n = 1, 2, \dots$, which provide approximation of order s, and for appropriately chosen sets Ω_n with $\#\Omega_n = O(n^{d-1})$, the sequence of spaces $Y_n, n = 1, 2, \dots$, given in (1.1), provide the following approximation: for $\lambda := s + (d-1)/2$,

$$E(f, Y_n)_{L_2(\mathbf{B}^d)} \le c(\lambda) n^{-\lambda} \|f\|_{W^\lambda(L_2(\mathbf{B}^d))}, \quad f \in W^\lambda(L_2(\mathbf{B}^d)). \tag{1.3}$$

Note that there is in a certain sense an unexpected gain in the multivariate approximation order $s + (d-1)/2$ over the univariate order s. This gain will be explained later (see section 9).

One can generate the space Y_n appearing in (1.3) by using very general univariate spaces X_n such as splines or wavelets. In particular, our results apply to feed-forward neural networks using a very general activation function σ. A complete discussion of the application to neural networks is given in section 9. In this introduction, we wish to illustrate the typical result by considering the following simple example. Let $\sigma = \chi_{[0,\infty)}$ and define X_n as the univariate space spanned by $\sigma(x - k/n)$, $0 \le k < n$. Then, defining Y_n for this X_n as described above, we obtain a space of dimension $O(n^d)$ of certain piecewise constant functions. The space Y_n can be realized computationally by a feed-forward neural network with $O(n^{d-1})$ computational nodes. In this case (see section 9 for details), (1.3) provides the approximation order $1 + \frac{d-1}{2}$. One might expect the estimate (1.3) to be 1 since we are using piecewise constants in the approximation. As noted in (1.3), the gain of $\frac{d-1}{2}$ in the approximation rate persists in general (see also Theorem 8.2).

There is a standard method in approximation theory (see [DL, Chapter 7]) which derives from (1.3) the estimate

$$E(f, Y_n)_{L_2(\mathbf{B}^d)} \le c \left(\omega_r(f, n^{-1})_{L_2(\mathbf{B}^d)} + \|f\|_{L_2(\mathbf{B}^d)} n^{-r} \right), \quad f \in L_2(\mathbf{B}^d) \tag{1.4}$$

with ω_r the rth order modulus of smoothness of f. In the case that Y_n contains all polynomials of total degree $< r$ (in d variables), the last term on the right can be eliminated.

Since Y_n is a linear space of dimension $O(n^d)$ then it follows from the general theory of n-widths that for all $m > 0$,

$$\sup_{\|f\|_{W^m(L_2(\mathbf{B}^d))} \le 1} E(f, Y_n) \ge c_0 n^{-m} \tag{1.5}$$

with $c_0 > 0$ a constant depending only on m and d. In this sense, the estimates (1.3) cannot be improved.

We also note that (1.3) shows that, in general, linear spaces of ridge functions are at least as efficient as other methods of multivariate approximation such as polynomials, wavelets, and splines.

This paper is an extension of the results from [DOP], where we considered the case $d = 2$. Throughout the paper we assume that $d > 2$, although most of the statements hold when $d = 2$.

The results of this paper differ from other work in this field in the following respects. We are able to begin with a very general class of univariate spaces X_n. Other authors (most notably Mhaskar and Micchelli [MM], [MM1], and Mhaskar [M]) have also considered approximation problems of the type treated here. The work of Mhaskar and Micchelli does not give the best order of approximation. Mhaskar [M] has given best possible results but only in the case that X_n is generated using a rather restrictive class of sigmoidal functions.

Our results are, for the present, limited to approximation in L_2, and it remains an important open question in ridge approximation to understand to what extent results such as those presented in this paper are valid in L_p, $p \neq 2$.

It is also an interesting question to understand which sets $\Omega_n \subset \mathbf{S}^{d-1}$, when used in defining the spaces Y_n, will provide the approximation order of (1.3). In the case $d = 2$, as was shown in [DOP], n equally spaced points on $\mathbf{S}^1$ are the most natural choice. There is no direct analogy of equally spaced points in $\mathbf{S}^{d-1}$, $d > 2$. It will become clear from section 4 that any set Ω_n which permits a cubature formula that is exact for spherical polynomials of degree $\leq n$ and with good localization properties will provide spaces Y_n which satisfy (1.3). Since we could not find in the literature examples of such sets Ω_n, we construct some in section 4. There should be more elegant and more natural constructions than ours. In some sense, one might expect that a natural quadrature formula might provide the analogue of equally spaced points in $\mathbf{S}^{d-1}$, $d > 2$.

We prove (1.3) by first understanding well the structure of ridge polynomials. Our main vehicle (given in section 3) is a fundamental orthogonal decomposition of a general function $f \in L_2(\mathbf{B}^d)$ into ridge polynomials. This decomposition uses the univariate Gegenbauer polynomials.

An outline of this paper is the following. The properties we need about Gegenbauer polynomials are given in section 2. In section 3, we give the fundamental orthogonal decomposition of functions in $L_2(\mathbf{B}^d)$ in terms of ridge polynomials. In section 4, we give our construction of cubature (quadrature) formulas. In sections 5–6, we introduce smoothness spaces (the Sobolev spaces) and recall their characterization by polynomial approximation. In section 7, we prove the main theorem about approximation by ridge functions. In section 8, we discuss how to improve the theorem of section 7 to be more amenable to applications. In section 9, we give some applications of our results, in particular to feed-forward neural networks.

Throughout the paper, the constants are denoted by $c, c_1, \ldots$ and they may vary at every occurrence. The constants usually depend on some parameters (like the dimension d) that will be sometimes indicated explicitly.

2. The Gegenbauer (ultraspherical) polynomials. Special functions appear naturally when we represent a general function in terms of ridge polynomials as will be done in the next section. In particular, the Gegenbauer polynomials will play an important role in this paper. In this section, we shall present the essential properties of Gegenbauer polynomials and bring out their role in the Radon transform. We refer the reader to [E] and [Sz] as general references for this section.

The Gegenbauer polynomials are usually defined by the following generating function

$$(1-2tz+z^2)^{-\lambda}=\sum_{m=0}^{\infty}C_m^\lambda(t)z^m,$$

where $|z|<1$, $|t|\le 1$, and $\lambda>0$. The coefficients $C_m^\lambda(t)$ are algebraic polynomials of degree m which are called the Gegenbauer polynomials associated with λ. The family of polynomials $\{C_m^\lambda\}_{m=0}^\infty$ is a complete orthogonal system for the weighted space $L_2(I,w)$, $I:=[-1,1]$, $w(t):=w_\lambda(t):=(1-t^2)^{\lambda-\frac12}$ and we have

$$\int_I C_m^\lambda(t)C_n^\lambda(t)w(t)\,dt=\begin{cases}0, & m\ne n\\ h_{n,\lambda}, & m=n\end{cases}\quad\text{with } h_{n,\lambda}:=\frac{\pi^{1/2}(2\lambda)_n\Gamma(\lambda+\frac12)}{(n+\lambda)n!\Gamma(\lambda)},\tag{2.1}$$

where we use here and later the standard notation

$$(a)_0:=0,\quad (a)_n:=a(a+1)\dots(a+n-1)=\Gamma(a+n)/\Gamma(a).$$

Also, we have

$$C_n^\lambda(-t)=(-1)^nC_n^\lambda(t),\quad C_n^\lambda(1)=\frac{(2\lambda)_n}{n!},\qquad\text{and}\quad C_0^\lambda(t)=1.\tag{2.2}$$

The Gegenbauer polynomials can also be defined by the following identity (called Rodrigues' formula):

$$C_n^\lambda(t)=(-1)^n\alpha_{n,\lambda}(1-t^2)^{-\lambda+\frac12}\left(\frac{d}{dt}\right)^n(1-t^2)^{n+\lambda-\frac12},\ \alpha_{n,\lambda}:=\frac{(2\lambda)_n}{n!2^n(\lambda+\frac12)_n}.\tag{2.3}$$

There is an identity that relates Gegenbauer polynomials with different weights:

$$\left(\frac{d}{dt}\right)^m C_n^\lambda(t)=2^m(\lambda)_mC_{n-m}^{\lambda+m}(t),\quad m=1,2,\dots,n.\tag{2.4}$$

Special cases of the Gegenbauer polynomials are the Legendre polynomials P_n and the Chebyshev polynomials of second kind U_n which correspond to $\lambda=1/2$ and $\lambda=1$, respectively. Namely,

$$P_n(t):=\frac{(-1)^n}{2^nn!}\left(\frac{d}{dt}\right)^n(1-t^2)^n=C_n^{1/2}(t),$$
$$U_n(t):=\frac{\sin(n+1)\arccos t}{\sqrt{1-t^2}}=C_n^1(t).$$

The Chebyshev polynomials of the first kind $T_n(t):=\cos n\arccos t$ can be considered as the Gegenbauer polynomials C_n^0 associated with the weight $w_0(t)=(1-t^2)^{-\frac12}$.

We shall also need the Gegenbauer polynomials C_n^λ when $\lambda<0$ and, in particular, when $\lambda=-1,-2,\dots$. Note that $\alpha_{n,\lambda}=0$ when $\lambda=-1,-2,\dots$ and $n>2\nu$. Therefore, we cannot use (2.3) to define $C_n^{-\nu}$ when $\nu=1,2,\dots$. However, we can define (see [Sz, Chapter IV])

$$C_n^\lambda(t):=\alpha(1-t^2)^{-\lambda+\frac12}\left(\frac{d}{dt}\right)^n(1-t^2)^{n+\lambda-\frac12},\quad\lambda<0,\tag{2.5}$$

where α is any constant independent of t. To our goals the normalization of C_n^λ ($\lambda<0$) is not essential. Identity (2.4) remains valid except for a constant factor (see [Sz, Chapter IV]): for any λ, we have

$$\left(\frac{d}{dt}\right)^m C_n^\lambda(t)=cC_{n-m}^{\lambda+m}(t),\quad m=1,2,\dots,n,\tag{2.6}$$

where c is independent of t.

The Gegenbauer polynomials play a fundamental role in inverting the Radon transform. We shall show in Lemma 2.1 that follows that the Gegenbauer polynomials C_n^λ for $\lambda = k$ and $\lambda = k + 1/2$ (k an integer) are eigenfunctions for certain differential operators that occur in the Radon transform inversion formula. These operators will play an important role in defining an equivalent norm for the weighted Sobolev spaces $W^s(L_2(I, w))$ (see section 8).

We begin with a brief discussion of the Hilbert transform H on $\mathbf{R}$ and its analogue $\mathbf{H}$ for the interval $I := [-1, 1]$. For any $g \in L_1(I)$ we define

$$\mathbf{H}g := Hg^\diamond \qquad \text{with} \qquad g^\diamond(t) := \begin{cases} g(t), & t \in I, \\ 0, & t \in (-\infty, \infty)\backslash I, \end{cases} \tag{2.7}$$

where $Hg^\diamond$ is the Hilbert transform of $g^\diamond$. It follows that

$$\mathbf{H}g(t) = \frac{1}{\pi}\text{p.v.} \int_{\mathbf{R}^1} \frac{g^\diamond(s)}{t-s}\, ds = \frac{1}{\pi}\text{p.v.} \int_I \frac{g(s)}{t-s}\, ds.$$

The analogue of the Hilbert transform on the circle $\mathbf{T}$ is the conjugate operator (see [Z, Chapter II]). If $g \in L_1(\mathbf{T})$, we denote its conjugate function by

$$\tilde{g}(\tau) := \frac{1}{2\pi}\text{p.v.} \int_{\mathbf{T}} g(\theta) \cot \frac{\tau - \theta}{2}\, d\theta.$$

For any (nonnegative) weight function w, let $L_2^0(I, w)$ be the space of all $g \in L_2(I, w)$ with weighted mean value zero: $\int_I g(t)\, w(t) dt = 0$. The following proposition gives some properties of $\mathbf{H}$ which we shall use.

PROPOSITION 2.1. *If $g \in L_1(I)$, we define $Tg(\theta) := \operatorname{sgn}\theta g(\cos\theta)\sin\theta$ for $\theta \in [-\pi, \pi)$. The Hilbert transform $\mathbf{H}$ satisfies the following properties:*

(a) *If $g \in L_1(I)$ then*

$$\mathbf{H}g(\cos\tau) = -\frac{1}{\sin\tau}\widetilde{Tg}(\tau) \quad \text{a.e. on } (0, \pi). \tag{2.8}$$

(b) *We have, on $(-1, 1)$, $\mathbf{H}w_1^{-1} = 0$,*

$$\mathbf{H}[w_1^{-1}T_{n+1}] = -U_n \quad \text{and} \quad \mathbf{H}[w_1 U_n] = T_{n+1} \quad \text{for } n = 0, 1, \ldots, \tag{2.9}$$

and hence

$$\mathbf{H}\frac{d}{dt}[w_1 U_n] = (n+1)U_n. \tag{2.10}$$

(c) *The functions $V_n := w_1^{-1}T_n$, $n = 0, 1, \ldots$ (in analogy to $\{U_n\}_{n=0}^\infty$) form a complete orthogonal system for $L_2(I, w_1)$.*

(d) *$\mathbf{H}$ is a one-to-one mapping of $L_2^0(I, w_1)$ onto $L_2(I, w_1)$ with*

$$\mathbf{H}^{-1}h = -\frac{1}{w_1}\mathbf{H}(w_1 h) \quad \text{for } h \in L_2(I, w_1)$$

and

$$\|\mathbf{H}g\|_{L_2(I,w_1)} = \|g\|_{L_2(I,w_1)} \quad \text{for } g \in L_2^0(I, w_1). \tag{2.11}$$

(e) *The operators* $\mathbf{H}$ *and* $\frac{d}{dt}$ *commute: for any polynomial* P*, we have*

$$\mathbf{H}\left(\frac{d}{dt}(w_1 P)\right) = \frac{d}{dt}\left(\mathbf{H}(w_1 P)\right).$$

Proof. (a) We apply the substitution $s = \cos\theta$ to the integral that defines $\mathbf{H}g$ and replace t by $\cos\tau$, $0 < \tau < \pi$ and obtain

$$\begin{aligned}\mathbf{H}g(\cos\tau) &= \frac{1}{\pi}\text{p.v.}\int_0^\pi \frac{Tg(\theta)}{\cos\tau - \cos\theta}\,d\theta \\ &= \frac{1}{2\pi}\text{p.v.}\int_{-\pi}^\pi \frac{Tg(\theta)}{\cos\tau - \cos\theta}\,d\theta,\end{aligned}$$

since the integrand is even. Note that $\text{p.v.}\int_I \ldots ds = \text{p.v.}\int_0^\pi \ldots d\theta$ above since the substituting function and its inverse are smooth enough. Now, we use the identity

$$\frac{1}{\cos\tau - \cos\theta} = -\frac{1}{2\sin\tau}\left(\cot\frac{\tau-\theta}{2} + \cot\frac{\tau+\theta}{2}\right)$$

to obtain

$$\mathbf{H}g(\cos\tau) = -\frac{1}{2\sin\tau}\left[\frac{1}{2\pi}\text{p.v.}\int_{-\pi}^\pi Tg(\theta)\cot\frac{\tau-\theta}{2}\,d\theta + \frac{1}{2\pi}\text{p.v.}\int_{-\pi}^\pi Tg(\theta)\cot\frac{\tau+\theta}{2}\,d\theta\right].$$

After substituting $\theta = -\theta'$ in the second integral above and using that Tg is even, we see that the two integrals are equal and, therefore, we obtain (a).

(b) For any function $g \in L_1(I, w_1^{-1})$, we have $T[w_1^{-1}g](\theta) = g(\cos\theta)$. Since the conjugate function of $\cos n\theta$ is $\sin n\theta$, $n = 0, 1, \ldots$, the first two statements in (b) follow from (a). Similar calculations give the last two statements.

(c) This is trivial.

(d) This follows from (b) by using the two bases for $L_2(I, w_1)$ given in (c).

(e) This follows from (2.10). □

We shall next show that the Gegenbauer polynomials are eigenfunctions of certain differential operators that arise in inverting the Radon transform. For functions g defined on $\mathbf{B}^d$, we introduce the following differential operators:

$$\Lambda g := \left(\frac{d}{dt}\right)^{d-1}\left[w_{d/2}g\right], \tag{2.12}$$

and

$$\mathcal{D} := \Lambda,\ d \text{ odd}, \quad \mathcal{D} := \mathbf{H}\Lambda,\ d \text{ even}. \tag{2.13}$$

LEMMA 2.1. *Let* $d \geq 2$ *and define* $\mathcal{U}_n := C_n^{d/2}$ *for* $n = 0, 1, \ldots$*. Then we have*

$$\mathcal{D}\mathcal{U}_n = (-1)^{[\frac{d-1}{2}]}\mu_n\mathcal{U}_n, \quad n = 0, 1, \ldots, \tag{2.14}$$

and

$$\Lambda^2\mathcal{U}_n = (-1)^{d-1}\mu_n^2\mathcal{U}_n, \quad n = 0, 1, \ldots, \tag{2.15}$$

where

$$\mu_n = (n+1)_{d-1} \asymp n^{d-1}, \quad n = 0, 1, \ldots. \tag{2.16}$$

Proof. We first consider (2.14) in the case when d is odd, $d = 2k+1$. From (2.3) and (2.4), we find

$$\begin{aligned}\mathcal{D}\mathcal{U}_n = \mathcal{D}C_n^{k+1/2} &= \left(\frac{d}{dt}\right)^{2k}\left[w_1^{2k}C_n^{k+1/2}\right] = c\left(\frac{d}{dt}\right)^{n+2k} w_1^{2n+2k}\\ &= c_1\left(\frac{d}{dt}\right)^{k} C_{n+k}^{1/2} = c_2 C_n^{k+1/2} = c_2\mathcal{U}_n.\end{aligned}$$

By examining the coefficients of t^n we obtain that $c_2 = (-1)^k\mu_n$. Thus (2.14) is proved in this case.

Assume now that d is even, $d =: 2k$. Then, again using (2.3), (2.4), and (2.10) (recall that $C_n^1 = U_n$) and the commutativity of $\frac{d}{dt}$ and $\mathbf{H}$, we obtain

$$\begin{aligned}\mathcal{D}\mathcal{U}_n = \mathcal{D}C_n^k &= \left(\frac{d}{dt}\right)^{2k-1}\mathbf{H}\left[w_1^{2k-1}C_n^k\right] = c\left(\frac{d}{dt}\right)^{2k-1}\mathbf{H}\left(\frac{d}{dt}\right)^{n} w_1^{2n+2k-1}\\ &= c\left(\frac{d}{dt}\right)^{k-1}\mathbf{H}\left(\frac{d}{dt}\right)^{n+k} w_1^{2n+2k-1} = c_1\left(\frac{d}{dt}\right)^{k-1}\mathbf{H}\frac{d}{dt}\left[w_1 C_{n+k-1}^1\right]\\ &= c_2\left(\frac{d}{dt}\right)^{k-1} C_{n+k-1}^1 = c_3 C_n^k = c_3 C_n^{d/2} = c_3\mathcal{U}_n.\end{aligned}$$

We can calculate c_3 as follows. Let $C_n^k(t) =: c_n t^n + \cdots$ and $U_r(t) =: a_r t^r + \cdots$ with $r := n+2k-2$. We find

$$\begin{aligned}&\left(\frac{d}{dt}\right)^{2k-1}\mathbf{H}\left[w_1^{2k-1}C_n^k(t)\right] = c_n\left(\frac{d}{dt}\right)^{2k-1}\mathbf{H}\left[w_1(t)\left((-1)^{k-1}t^{n+2k-2}+\cdots\right)\right]\\ &= (-1)^{k-1}\frac{c_n}{a_r}\left(\frac{d}{dt}\right)^{2k-2}\left[\mathbf{H}\frac{d}{dt}\left[w_1(t)U_{n+k-2}(t)+\cdots\right]\right]\\ &= (-1)^{k-1}\frac{c_n}{a_r}(n+2k-1)\left(\frac{d}{dt}\right)^{2k-2}\left[U_{n+2k-2}(t)+\cdots\right]\\ &= (-1)^{k-1}c_n(n+2k-1)\left(\frac{d}{dt}\right)^{2k-2}\left(t^{n+2k-2}+\cdots\right)\\ &= (-1)^{k-1}(n+1)_{2k-1}c_n\left(t^n+\cdots\right) = (-1)^{k-1}(n+1)_{2k-1}C_n^k(t),\end{aligned}$$

where we used identities (2.3), (2.4), and (2.10). Thus (2.14) is proved in this case as well.

Finally, we consider (2.15). From (2.3) and (2.5), respectively, we have

$$\begin{aligned}\Lambda\mathcal{U}_n = \Lambda C_n^{d/2} &= c\left(\frac{d}{dt}\right)^{n+d-1}\left[(1-t^2)^{n+d/2-1/2}\right]\\ &= c\left(\frac{d}{dt}\right)^{n+d-1}\left[(1-t^2)^{n+d-1}(1-t^2)^{-d/2+1/2}\right]\\ &= c_1(1-t^2)^{-d/2+1/2}C_{n+d-1}^{-d/2+1}.\end{aligned}$$

Hence, applying Λ once again and using (2.6) gives

$$\Lambda^2\mathcal{U}_n = \Lambda^2 C_n^{d/2} = c_1\left(\frac{d}{dt}\right)^{d-1} C_{n+d-1}^{-d/2+1} = c_2 C_n^{d/2} = c_2\mathcal{U}_n.$$

By calculating the coefficients of t^n, we find that $c_2 = (-1)^{d-1}\mu_n^2$ and we arrive at (2.15). ∎

3. An orthogonal decomposition of $L_2(\mathbf{B}^d)$ in terms of ridge polynomials. Since we are interested in approximating functions $f \in L_2(\mathbf{B}^d)$ by elements from spaces of ridge functions, it is natural to find a decomposition of f in terms of fundamental building blocks of ridge functions. We shall show in this section that we can take as the building blocks certain ridge polynomials. We begin by describing this decomposition.

If $f, g \in L_2(\mathbf{B}^d)$, we define the inner product

$$\langle f, g\rangle := \int_{\mathbf{B}^d} f(\mathbf{x})\overline{g(\mathbf{x})}\,d\mathbf{x}. \tag{3.1}$$

This inner product induces the norm

$$\|f\|_{L_2(\mathbf{B}^d)} := \left(\int_{\mathbf{B}^d} |f(\mathbf{x})|^2\,d\mathbf{x}\right)^{1/2}.$$

We also define, for $f, g \in L_2(\mathbf{S}^{d-1})$, the inner product

$$(f, g) := \int_{\mathbf{S}^{d-1}} f(\xi)\overline{g(\xi)}\,d\xi \tag{3.2}$$

and the norm

$$\|f\|_{L_2(\mathbf{S}^{d-1})} := \left(\int_{\mathbf{S}^{d-1}} |f(\xi)|^2\,d\xi\right)^{1/2},$$

where $d\xi$ stands for the area (volume) element on $\mathbf{S}^{d-1}$ the unit sphere in $\mathbf{R}^d$.

The Gegenbauer polynomials $C_n^{d/2}$ are the building blocks for our decomposition. Let

$$\mathcal{U}_n := \left(h_{n,d/2}\right)^{-1/2} C_n^{d/2}, \quad n = 0, 1, \dots, \tag{3.3}$$

where $h_{n,d/2}$ is from (2.1). Then $\|\mathcal{U}_n\|_{L_2(I,w)} = 1$ and hence $\{\mathcal{U}_n\}_{n=0}^\infty$ is a complete orthonormal system for the weighted space $L_2(I, w)$, $w(t) := w_{d/2}(t) = (1-t^2)^{\frac{d-1}{2}}$. Of course, $\mathcal{U}_n$ depends on the space dimension d, but we are suppressing this dependence in our notation. The reader should think of the space dimension d as arbitrary but fixed throughout.

Let $\mathcal{P}_n$ denote the set of all algebraic polynomials of total degree n in d real variables. That is, each $P \in \mathcal{P}_n$ is a linear combination of monomials $\mathbf{x}^{\mathbf{m}} := x_1^{m_1} \dots x_d^{m_d}$ with $\mathbf{x} := (x_1, \dots, x_d)$, $\mathbf{m}$ is a d-tuple $(m_1, \dots, m_d)$ of nonnegative integers, and $|\mathbf{m}| := m_1 + \cdots + m_d \le n$.

The polynomials $\mathcal{U}_n(\xi \cdot \mathbf{x})$, $\xi \in \mathbf{S}^{d-1}$, are in $\mathcal{P}_n$ and $\mathcal{U}_n(\xi \cdot \mathbf{x})$ are orthogonal to $\mathcal{P}_{n-1}$ in $L_2(\mathbf{B}^d)$ (proved in the appendix):

$$\int_{\mathbf{B}^d} \mathcal{U}_n(\xi \cdot \mathbf{x})P(\mathbf{x})\,d\mathbf{x} = 0 \quad \text{for } \xi \in \mathbf{S}^{d-1} \text{ and } P \in \mathcal{P}_{n-1}. \tag{3.4}$$

This is why the ridge polynomials $\mathcal{U}_n(\xi \cdot \mathbf{x})$ occur in our decomposition of $L_2(\mathbf{B}^d)$.

THEOREM 3.1. *Each function $f \in L_2(\mathbf{B}^d)$ can be represented uniquely as*

$$f \overset{L_2}{=} \sum_{n=0}^\infty Q_n(f), \tag{3.5}$$

where

$$Q_n(f,\mathbf{x}) := \nu_n \int_{\mathbf{S}^{d-1}} A_n(f,\xi)\mathcal{U}_n(\xi\cdot\mathbf{x})\,d\xi \tag{3.6}$$

with

$$A_n(f,\xi) := \int_{\mathbf{B}^d} f(\mathbf{y})\mathcal{U}_n(\xi\cdot\mathbf{y})\,d\mathbf{y}, \tag{3.7}$$

and

$$\nu_n := \frac{(n+1)_{d-1}}{2(2\pi)^{d-1}} = \frac{(n+1)(n+2)\cdots(n+d-1)}{2(2\pi)^{d-1}}. \tag{3.8}$$

Moreover, the operators Q_n, $n = 0,1,\ldots$, are the orthogonal projectors from $L_2(\mathbf{B}^d)$ onto $\mathcal{P}_n \ominus \mathcal{P}_{n-1}$ and the following Parseval identity holds

$$\|f\|^2_{L_2(\mathbf{B}^d)} = \sum_{n=0}^{\infty} \|Q_n(f)\|^2_{L_2(\mathbf{B}^d)} = \sum_{n=0}^{\infty} \nu_n \|A_n\|^2_{L_2(\mathbf{S}^{d-1})}. \tag{3.9}$$

Next we make a few remarks which will help explain the nature of this decomposition.

(i) For each $n = 0,1,\ldots$, the function $Q_n(f)$ is an algebraic polynomial (in d variables) of degree n. Indeed, each of the $\mathcal{U}_n(\xi\cdot\mathbf{x})$ is a ridge polynomial of degree n and $Q_n(f)$ is a linear combination of these.

(ii) For each $n = 0,1,\ldots$, the function $A_n(f,\xi)$, $\xi \in \mathbf{S}^{d-1}$, is a spherical polynomial of degree n. This follows from the fact that each of the $\mathcal{U}_n(\xi\cdot\mathbf{x})$, $x \in \mathbf{B}^d$, is of this type.

(iii) The constants ν_n, $n = 0,1,\ldots$, are eigenvalues which occur in the Radon inversion formula (see (3.26)).

(iv) Among other reasons, the polynomials $\mathcal{U}_n$ occur in this formula because for each $\xi \in \mathbf{S}^{d-1}$ the weight $w_{d/2}(t) = (1-t^2)^{(d-1)/2}$ is a constant multiple of the $d-1$-dimensional volume of the intersection of $\mathbf{B}^d$ with the hyperplane $\mathbf{x}\cdot\xi = t$.

(v) The orthogonality of the functions $Q_n(f)$ occurs because for each $\xi \in \mathbf{S}^{d-1}$, the polynomial $\mathcal{U}_n(\mathbf{x}\cdot\xi)$ is orthogonal to all algebraic polynomials of degree $< n$ on $\mathbf{B}^d$ (see (3.4)).

(vi) One can imagine that the integral representation of $Q_n(f)$ can be rewritten as a discrete sum by using some sort of quadrature formula on $\mathbf{S}^{d-1}$ and thereby obtain a discrete decomposition of f in terms of ridge polynomials. In the case $d = 2$, one can simply take the canonical quadrature formula for integrating spherical polynomials (i.e., trigonometric polynomials) which uses equally spaced points on the unit circle. This then gives the orthonormal system $\{\mathcal{U}_n(\omega\cdot\mathbf{x})\}$, $\omega \in \Omega_n$, $n = 0,1,\ldots$, where $\Omega_n := \{(\cos k\pi/n, \sin k\pi/n)\}_{k=1}^n$. This was used in [DOP] as the vehicle for proving approximation results for ridge functions in two variables. In the case $d \geq 3$, we know no analogous quadrature formula. This necessitates a substantial effort (executed in the following section) to derive (less elegant) quadrature formulas which can be used to discretize the integral representation of $Q_n(f)$.

(vii) The decomposition of Theorem 3.1 is in essence known (see, e.g., [RK] and [LS] for the case $d = 2$). However, we could find no reference which gives it in the above form.

There are several ways in which the decomposition of Theorem 3.1 can be derived. One approach is to derive it from the theory of spherical harmonics. A second approach is Radon transforms and in particular (3.5) is a rewriting of the Radon inversion formula (see [RK]). We shall briefly explain this at the end of this section.

We shall give a simple and direct proof of this decomposition using fundamental identities for the ridge polynomials $\mathcal{U}_n(\xi \cdot \mathbf{x})$, $\xi \in \mathbf{S}^{d-1}$. To keep our exposition more fluid, we shall state these identities without proof and relegate the proofs to the appendix.

We start with the following two fundamental identities (proved in the appendix): for each $\xi, \eta \in \mathbf{S}^{d-1}$, we have

$$\int_{\mathbf{B}^d} \mathcal{U}_n(\xi \cdot \mathbf{x})\mathcal{U}_n(\eta \cdot \mathbf{x})\, d\mathbf{x} = \frac{\mathcal{U}_n(\xi \cdot \eta)}{\mathcal{U}_n(1)}, \tag{3.10}$$

and, for each $\eta \in \mathbf{S}^{d-1}$, we have

$$\int_{\mathbf{S}^{d-1}} \mathcal{U}_n(\xi \cdot \mathbf{x})\mathcal{U}_n(\xi \cdot \eta)\, d\xi = \frac{\mathcal{U}_n(1)}{\nu_n}\mathcal{U}_n(\eta \cdot \mathbf{x}). \tag{3.11}$$

Proof of Theorem 3.1. Let $f \in L_2(\mathbf{B}^d)$, $d > 2$. From remark (i) and (3.4), it follows that $Q_n(f)$ is in $\mathcal{P}_n \ominus \mathcal{P}_{n-1}$. From identities (3.10) and (3.11), we have $Q_n(g) = g$ whenever $g(\mathbf{x}) = \mathcal{U}_n(\eta \cdot \mathbf{x})$, $\eta \in \mathbf{S}^{d-1}$. Therefore, $Q_n^2 = Q_n$ and hence Q_n is a projector onto a subspace Y_n of $\mathcal{P}_n \ominus \mathcal{P}_{n-1}$. Thus, to prove (3.5), it remains only to show that

$$\dim(Y_n) = \dim(\mathcal{P}_n \ominus \mathcal{P}_{n-1}) = \dim(\mathcal{P}_n^h), \tag{3.12}$$

where $\mathcal{P}_n^h$ denotes the space of all homogeneous polynomials of degree n.

To prove (3.12), we recall a few well-known facts about spherical harmonics which can be found in Stein and Weiss [SW, Chapter 4]; see also [Se]. Let $\mathcal{H}_n$ denote the space of spherical harmonics of degree n; i.e., $\mathcal{H}_n$ is the set of those functions on $\mathbf{S}^{d-1}$ which are the restriction to $\mathbf{S}^{d-1}$ of a function from $\mathcal{P}_n^h$ which is harmonic in $\mathbf{B}^d$. The spherical harmonics of degree n are orthogonal to those of dimension $m \neq n$ with respect to the inner product (3.2). We have

$$\dim(\mathcal{H}_n) = N(d,n) := \left(\begin{array}{c} n+d-1 \\ n \end{array} \right) - \left(\begin{array}{c} n+d-3 \\ n-2 \end{array} \right) \tag{3.13}$$

and

$$\dim(\mathcal{P}_n^h) = \dim(\mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon), \tag{3.14}$$

where $\epsilon = 0$ if n is even and $\epsilon = 1$ if n is odd.

Write

$$K_n(t) := \frac{N(d,n)}{|\mathbf{S}^{d-1}| C_n^{(d-2)/2}(1)} C_n^{(d-2)/2}(t),$$

where $|\mathbf{S}^{d-1}| := \int_{\mathbf{S}^{d-1}} 1\, d\xi = \frac{2\pi^{d/2}}{\Gamma(d/2)}$ is the surface area of $\mathbf{S}^{d-1}$. The function $K_n(\xi \cdot \eta)$ is the reproducing kernel for $\mathcal{H}_n$; i.e.,

$$\int_{\mathbf{S}^{d-1}} S(\xi) K_n(\xi \cdot \eta) d\xi = S(\eta), \quad S \in \mathcal{H}_n. \tag{3.15}$$

Moreover, a simple identity for Gegenbauer polynomials (see Appendix A3) gives that

$$K_n + K_{n-2} + \cdots + K_\epsilon = \frac{C_n^{d/2}}{|\mathbf{S}^{d-1}|} = \frac{\nu_n \mathcal{U}_n}{\mathcal{U}_n(1)}. \tag{3.16}$$

Hence, the right side of (3.16) is the reproducing kernel for $\mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$; i.e.,

$$\int_{\mathbf{S}^{d-1}} S(\xi) \frac{\nu_n}{\mathcal{U}_n(1)} \mathcal{U}_n(\xi \cdot \eta) d\xi = S(\eta), \quad S \in \mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon. \tag{3.17}$$

Note that $A_n(f, \xi)$ is a spherical polynomial of degree n since $\mathcal{U}_n(\xi \cdot \mathbf{y})$ is a spherical polynomial of degree n in ξ. We have $\mathcal{U}_n(-t) = (-1)^n \mathcal{U}_n(t)$ (see (2.2)) and hence $A_n(f, -\xi) = (-1)^n A_n(f, \xi)$. Therefore, $A_n(f) \in \mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$. Thus, Q_n can be considered as a linear operator mapping $\mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$ into Y_n. On the other hand, after multiplying both sides of (3.6) by $\mathcal{U}_n(\eta \cdot \mathbf{x})$ and integrating over $\mathbf{B}^d$ we obtain

$$\begin{aligned} \int_{\mathbf{B}^d} Q_n(f, \mathbf{x}) \mathcal{U}_n(\eta \cdot \mathbf{x}) \, d\mathbf{x} &= \int_{\mathbf{S}^{d-1}} A_n(f, \xi) \left(\nu_n \int_{\mathbf{B}^d} \mathcal{U}_n(\eta \cdot \mathbf{x}) \mathcal{U}_n(\xi \cdot \mathbf{x}) \, d\mathbf{x} \right) d\xi \\ &= \int_{\mathbf{S}^{d-1}} A_n(f, \xi) \frac{\nu_n}{\mathcal{U}_n(1)} \mathcal{U}_n(\eta \cdot \xi) \, d\xi = A_n(f, \xi), \end{aligned}$$

where we used (3.10) and (3.17). Hence, A_n is an operator mapping Y_n onto $\mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$ and it is the inverse operator of Q_n. Therefore, $\dim(Y_n) = \dim(\mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon)$ which together with (3.14) implies (3.12).

Since $Q_n(f)$ is in $\mathcal{P}_n \ominus \mathcal{P}_{n-1}$, it is orthogonal to $Q_j(f)$, $j \neq n$, and therefore we have the first equality in (3.9). For the proof of the second equality in (3.9), we use (3.10) to write

$$\begin{aligned} \int_{\mathbf{B}^d} Q_n(f, \mathbf{x})^2 \, d\mathbf{x} &= \nu_n^2 \int_{\mathbf{S}^{d-1}} \int_{\mathbf{S}^{d-1}} \int_{\mathbf{B}^d} A_n(f, \xi) A_n(f, \eta) \mathcal{U}_n(\xi \cdot \mathbf{x}) U_n(\eta \cdot \mathbf{x}) \, d\mathbf{x} \, d\xi \, d\eta \\ &= \nu_n^2 \int_{\mathbf{S}^{d-1}} \int_{\mathbf{S}^{d-1}} A_n(f, \xi) A_n(f, \eta) \frac{\mathcal{U}_n(\xi \cdot \eta)}{\mathcal{U}_n(1)} d\xi \, d\eta. \end{aligned}$$

Since $A_n(f) \in \mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$, then we can use (3.17) to complete the integral with respect to η above. We get

$$\int_{\mathbf{B}^d} Q_n(f, \mathbf{x})^2 \, d\mathbf{x} = \nu_n \int_{\mathbf{S}^{d-1}} A_n(f, \xi)^2 \, d\xi.$$

This completes the proof of (3.9) and the theorem. □

In the same way that we have proved (3.9) of Theorem 3.1 we obtain the following formulas for calculating inner products:

$$\langle f, g \rangle = \sum_{n=0}^{\infty} \nu_n \int_{\mathbf{S}^{d-1}} A_n(f, \xi) A_n(g, \xi) \, d\xi. \tag{3.18}$$

We next consider the decomposition (3.5) for ridge functions. Let r be a univariate function in $L_2(I, w)$, $w := w_{d/2}$. Then

$$r(t) = \sum_{n=0}^{\infty} \hat{r}(n) \mathcal{U}_n(t), \quad \hat{r}(n) := \int_I r(t) \mathcal{U}_n(t) w(t) \, dt. \tag{3.19}$$

It follows that for any $\rho \in \mathbf{S}^{d-1}$, the ridge function $R(\mathbf{x}) := r(\rho \cdot \mathbf{x})$ has the representation

$$R(\mathbf{x}) = \sum_{n=0}^{\infty} \hat{r}(n) \mathcal{U}_n(\rho \cdot \mathbf{x}). \tag{3.20}$$

Using (3.10) and (3.4), we see that

$$A_n(R, \xi) = \hat{r}(n) \frac{\mathcal{U}_n(\rho \cdot \xi)}{\mathcal{U}_n(1)}. \tag{3.21}$$

Moreover, if R_1 and R_2 are two such ridge functions corresponding to r_1, ρ_1 and r_2, ρ_2, respectively, then from (3.4), (3.11), and (3.18) we have

$$\langle R_1, R_2 \rangle = \sum_{n=0}^{\infty} \hat{r}_1(n) \hat{r}_2(n) \frac{\mathcal{U}_n(\rho_1 \cdot \rho_2)}{\mathcal{U}_n(1)}. \tag{3.22}$$

There is another approach to deducing the decomposition of Theorem 3.1 which we want to mention since it brings out the connections between this paper and Radon transforms. For each $f \in L_1(\mathbf{B}^d)$ the Radon transform is defined by

$$\mathcal{R}(f; \xi, t) := \int_{\xi^\perp \bigcap \mathbf{B}^d} f(t\xi + \mathbf{y})\, d\mathbf{y}, \tag{3.23}$$

where $\xi \in \mathbf{S}^{d-1}$, $t \in [-1, 1]$, and $\xi^\perp := \{\mathbf{y} \in \mathbf{R}^d : \mathbf{y} \cdot \xi = 0\}$. So, the integration is over the intersection of the hyperplane $\mathbf{y} \cdot \xi = t$ and $\mathbf{B}^d$.

We can recover f from its Radon transform by using the Radon transform inversion formula. The Radon transform inversion formula uses the operator (see, e.g., [L])

$$Kg(t) := K_t g(t) := \begin{cases} \frac{(-1)^{\frac{d-1}{2}}}{2(2\pi)^{d-1}} \left(\frac{d}{dt}\right)^{d-1} g(t) & \text{for } d \text{ odd}, \\ \\ \frac{(-1)^{\frac{d}{2}-1}}{2(2\pi)^{d-1}} \mathbf{H} \left(\frac{d}{dt}\right)^{d-1} g(t) & \text{for } d \text{ even}, \end{cases} \tag{3.24}$$

where $\mathbf{H}$ is the Hilbert transform (see (2.7)). The following relation is the Radon inversion formula for functions defined on $\mathbf{B}^d$: for every sufficiently smooth function f supported on $\mathbf{B}^d$

$$f(\mathbf{x}) = \int_{\mathbf{S}^{d-1}} h(\xi, \mathbf{x} \cdot \xi)\, d\xi \quad \text{with} \quad h(\xi, t) := K_t \mathcal{R}(f; \xi, t). \tag{3.25}$$

Lemma 2.1 gives that the functions $\mathcal{U}_n$ are eigenfunctions for the operator $K(w\bullet)$:

$$K(w\mathcal{U}_n) = \nu_n \mathcal{U}_n. \tag{3.26}$$

We now show the idea of using the Radon inversion formula to derive a representation of f in terms of the ridge polynomials $\{\mathcal{U}_n(\mathbf{x} \cdot \xi)\}$. Since $\{\mathcal{U}_n\}_{n=0}^{\infty}$ is a complete orthonormal system for $L_2(I, w)$, we can expand $\mathcal{R}(f; \xi, \bullet)/w$ in terms of the $\{\mathcal{U}_n\}_{n=0}^{\infty}$ to obtain

$$\frac{\mathcal{R}(f; \xi, \bullet)}{w} = \sum_{n=0}^{\infty} A_n(\xi)\, \mathcal{U}_n, \tag{3.27}$$

with

$$A_n(\xi) := \int_I \mathcal{R}(f;\xi,t)\,\mathcal{U}_n(t)\,dt = \int_{\mathbf{B}^d} f(\mathbf{y})\,\mathcal{U}_n(\mathbf{y}\cdot\xi)\,d\mathbf{y}. \tag{3.28}$$

After multiplying both sides of (3.27) by the weight w and applying the operator $K := K_t$ we get

$$K\mathcal{R}(f;\xi,\bullet) = \sum_{n=0}^{\infty} A_n(\xi) K\,[w\,\mathcal{U}_n] = \sum_{n=0}^{\infty} \nu_n A_n(\xi)\,\mathcal{U}_n,$$

where we used (3.26). Finally, we insert the above in (3.25) and find

$$f(\mathbf{x}) = \int_{\mathbf{S}^{d-1}} K\mathcal{R}(f;\xi,\mathbf{x}\cdot\xi)\,d\xi = \sum_{n=0}^{\infty} \nu_n \int_{\mathbf{S}^{d-1}} A_n(\xi)\mathcal{U}_n(\mathbf{x}\cdot\xi)\,d\xi$$

which is the decomposition from Theorem 3.1. We leave the details of verifying this approach to the reader.

4. Discrete representation of functions and norms. In this section we shall deduce from Theorem 3.1 a discrete representation of functions by ridge polynomials. To this end we shall use a cubature formula for integration on $\mathbf{S}^{d-1}$, $d > 2$. We need a cubature formula that is exact for all spherical polynomials of degree n. In the case $d = 2$ we used in [DOP] a quadrature formula with equally spaced nodes on the unit circle. Unfortunately, we do not know any "equally spaced points" on $\mathbf{S}^{d-1}$, $d > 2$. Also, we do not know effectively any cubature formula with near equally spaced nodes on $\mathbf{S}^{d-1}$. For this reason we shall use a cubature formula, determined by using spherical coordinates on $\mathbf{S}^{d-1}$. The results of this section are somewhat technical and the reader may just wish to read them briefly at first and proceed to section 5.

The spherical coordinates $(\theta,\phi) := (\theta_1,\theta_2,\dots,\theta_{d-2},\phi)$ on $\mathbf{S}^{d-1}$ are defined as usual by

$$\xi_1 = \cos\theta_1,\ \ \xi_2 = \sin\theta_1\cos\theta_2,\ \ \dots,\ \ \xi_{d-2} = \sin\theta_1\sin\theta_2\dots\sin\theta_{d-3}\cos\theta_{d-2},$$
$$\xi_{d-1} = \sin\theta_1\sin\theta_2\dots\sin\theta_{d-3}\sin\theta_{d-2}\cos\phi,\ \ \xi_d = \sin\theta_1\sin\theta_2\dots\sin\theta_{d-3}\sin\theta_{d-2}\sin\phi,$$

$0 \le \theta_j \le \pi$, $j = 1,2,\dots,d-2$; $0 \le \phi < 2\pi$. We shall denote these identities in vector form briefly by $\xi := \xi(\theta,\phi)$. In these coordinates, the surface element $d\xi$ of $\mathbf{S}^{d-1}$ becomes

$$d\xi = (\sin\theta_1)^{d-2}(\sin\theta_2)^{d-3}\dots\sin\theta_{d-2}\,d\theta_1\,d\theta_2\dots\,d\theta_{d-2}\,d\phi =: J(\theta)\,d\theta\,d\phi. \tag{4.1}$$

We have the following identity for integration in spherical coordinates

$$\int_{\mathbf{S}^{d-1}} f(\xi)\,d\xi = \int_0^{\pi}\dots\int_0^{\pi}\int_0^{2\pi} f(\xi(\theta,\phi))J(\theta)\,d\theta_1\dots\,d\theta_{d-2}\,d\phi, \tag{4.2}$$

where $J(\theta)$ is the Jacobian given by (4.1). We shall use this to define our cubature.

We wish to construct a cubature that is exact for all spherical polynomials of degree $2n$. Every spherical polynomial of degree $2n$ can obviously be represented in spherical coordinates as a linear combination of terms

$$(\cos\phi)^{k_{d-1}}(\sin\phi)^{\ell_{d-1}}\prod_{j=1}^{d-2}(\cos\theta_j)^{k_j}(\sin\theta_j)^{\ell_j}, \tag{4.3}$$

where $k_j, \ell_j \geq 0$, and $\max\{k_j+\ell_j : j = 1, 2, \ldots, d-1\} \leq 2(d-1)n$. Also, the Jacobian J is represented in the same terms (see (4.1)). So, we need quadrature formulas for integration over $[0, 2\pi]$ and $[0, \pi]$ that are exact for trigonometric polynomial of degree $2(d-1)n + d - 2$.

We shall use the following quadrature formula for integration on $[0, 2\pi]$ with respect to ϕ

$$Q_{\phi,k}(g) := \sum_{j=0}^{2k} \varrho_j g(\gamma_j) \quad \sim \quad \int_0^{2\pi} g(\phi)\, d\phi, \tag{4.4}$$

where $\gamma_j := \frac{\pi}{2k+1} + \frac{2\pi j}{2k+1}$ and $\varrho_j := \frac{2\pi}{2k+1}$. The quadrature (4.4) is exact for all trigonometric polynomials of degree k (see [Z, Chapter X]).

Since $0 \leq \theta_j \leq \pi$, we need a quadrature for integration over $[0, \pi]$ that is exact for all trigonometric polynomials of degree k. In addition to this, the quadrature should have good localization properties. We also need to control (asymptotically) the nodes and the coefficients of the quadrature. Since we do not know any quadrature like this, we shall construct one in the following lemma.

LEMMA 4.1. *For any $k = 1, 2, \ldots$, there exists a quadrature*

$$Q_{\theta,k}(g) = \sum_{j=0}^{2k} \lambda_j g(\beta_j) \quad \sim \quad \int_0^{\pi} g(\theta)\, d\theta \tag{4.5}$$

with the following properties:

(a) *$Q_{\theta,k}(g)$ is exact for all trigonometric polynomials of degree k;*

(b) $0 < \beta_0 < \beta_1 < \cdots < \beta_{2k} < \pi$,

$$\beta_j - \beta_{j-1} \leq \pi k^{-1}, \quad j = 0, 1, \ldots, 2k+1; \tag{4.6}$$

(c)

$$0 < \lambda_j \leq c\left(\beta_{j+1} - \beta_{j-1}\right), \quad j = 0, 1, \ldots, 2k, \tag{4.7}$$

where $\beta_{-1} := 0$, $\beta_{2k+1} := \pi$, and c is an absolute constant.

The exact values of the nodes β_j and the coefficients λ_j of the quadrature (4.5) *are given in Remark* 4.1 *below.*

Proof. For symmetry reasons we shall prove the lemma with the interval of integration $[0, \pi]$ replaced by $[-\pi/2, \pi/2]$. We shall build a quadrature

$$Q_k(g) = \sum_{j=-k}^{k} \lambda_j g(\theta_j) \quad \sim \quad \int_{-\pi/2}^{\pi/2} g(\theta)\, d\theta \tag{4.8}$$

with symmetric nodes θ_j and coefficients λ_j ($\theta_{-j} = \theta_j$, $\theta_0 := 0$, and $\lambda_{-j} = \lambda_j$). Then $Q_k(g)$ will be automatically exact for odd polynomials. Therefore, it is enough to construct $Q_k(g)$ exact only for all even trigonometric polynomials of degree k. To this end it is sufficient to have

$$Q_k(P(\cos\theta)) := \sum_{j=-k}^{k} \lambda_j P(\cos\theta_j) = \int_{-\pi/2}^{\pi/2} P(\cos\theta)\, d\theta = 2\int_0^{\pi/2} P(\cos\theta)\, d\theta \tag{4.9}$$

for each algebraic polynomial P of degree k. We shall apply the substitution $\theta := \theta(\alpha) := \arccos\left(\cos^2\frac{\alpha}{2}\right)$ to the last integral in (4.9). Simple calculations show that

$$\Delta(\alpha) := \frac{d}{d\alpha}\theta(\alpha) = \frac{\cos\frac{\alpha}{2}}{\sqrt{1+\cos^2\frac{\alpha}{2}}} \tag{4.10}$$

and hence $\theta(\alpha)$ is increasing on $[0,\pi]$ and maps $[0,\pi]$ on $[0,\pi/2]$. We obtain

$$\int_0^{\pi/2} P(\cos\theta)\,d\theta = \int_0^{\pi} P\left(\cos^2\frac{\alpha}{2}\right)\Delta(\alpha)\,d\alpha = \frac{1}{2}\int_{-\pi}^{\pi} P\left(\cos^2\frac{\alpha}{2}\right)\Delta(\alpha)\,d\alpha, \tag{4.11}$$

where we used that the integrand is even. We now extend $\Delta(\alpha)$ 2π-periodically by $\Delta(\alpha) := |\cos\frac{\alpha}{2}| \Big/ \sqrt{1+\cos^2\frac{\alpha}{2}}$.

We shall use the Dirichlet kernel $D_k(u) := \frac{\sin(k+1/2)u}{2\sin u/2}$ to interpolate the trigonometric polynomial of degree m: $P\left(\cos^2\frac{\alpha}{2}\right) = P\left(\frac{1+\cos\alpha}{2}\right)$ at the points $\alpha_j := \frac{2\pi j}{2k+1}$, $j = 0, \pm1, \dots, \pm k$. We have (see [Z, Chapter X])

$$P\left(\cos^2\frac{\alpha}{2}\right) = \frac{2}{2k+1}\sum_{j=-k}^{k} P\left(\cos^2\frac{\alpha_j}{2}\right) D_k(\alpha-\alpha_j).$$

This and (4.11) imply

$$\begin{aligned}\int_0^{\pi/2} P(\cos\alpha)\,d\alpha &= \frac{1}{2k+1}\sum_{j=-k}^{k} P\left(\cos^2\frac{\alpha_j}{2}\right)\int_{-\pi}^{\pi}\Delta(\alpha)D_k(\alpha-\alpha_j)\,d\alpha \\ &= \sum_{j=-k}^{k}\eta_j P\left(\cos^2\frac{\alpha_j}{2}\right) \qquad (4.12)\\ &= \eta_0 P\left(\cos^2\frac{\alpha_0}{2}\right) + \sum_{j=1}^{k} 2\eta_j P\left(\cos^2\frac{\alpha_j}{2}\right),\end{aligned}$$

where

$$\eta_j := \frac{1}{2k+1}\int_{-\pi}^{\pi}\Delta(\alpha)D_k(\alpha-\alpha_j)\,d\alpha \tag{4.13}$$

and we used that $\eta_{-j} = \eta_j$ since Δ is even and $\alpha_{-j} = -\alpha_j$.

We now define the nodes and the coefficients of our quadrature. Set

$$\theta_j := \arccos\left(\cos^2\frac{\alpha_j}{2}\right) \text{ for } j = 0,1,\dots,k \text{ and } \theta_j := -\theta_{-j} \text{ for } j = -1,-2,\dots,-k.$$

Also, set

$$\lambda_j := 2\eta_j \quad \text{for} \quad j = 0,1,\dots,k \quad \text{and} \quad \lambda_j := \lambda_{-j} \quad \text{for} \quad j = -1,-2,\dots,-k.$$

We obtain by (4.9), (4.12) and the symmetry that quadrature (4.8) with the above defined nodes and coefficients is exact for all trigonometric polynomials of degree m. It remains to prove that the nodes and the coefficients of the quadrature satisfy the required properties.

We have $\theta_j - \theta_{j-1} = \theta'(\zeta_j)(\alpha_j - \alpha_{j-1}) = \Delta(\zeta_j)(\alpha_j - \alpha_{j-1})$ for some $\zeta_j \in (\alpha_{j-1}, \alpha_j)$ and hence, by (4.10),

$$\frac{1}{\sqrt{2}}\left(\cos\frac{\alpha_j}{2}\right)\frac{2\pi}{2k+1} \le \theta_j - \theta_{j-1} \le \left(\cos\frac{\alpha_j}{2}\right)\frac{2\pi}{2k+1} < \frac{\pi}{k}, \quad j = 1, 2, \dots, k. \tag{4.14}$$

Therefore, the proof will be completed if we show that

$$0 < \eta_j \le ck^{-1}\cos\frac{\alpha_j}{2}, \quad j = 1, 2, \dots, k. \tag{4.15}$$

By (4.13) it follows that

$$\eta_j = \pi(2k+1)^{-1}S_k(\Delta)(\alpha_j), \quad \text{where} \quad S_k(\Delta)(\alpha) := \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi}\Delta(\beta)D_k(\beta-\alpha)\,d\beta$$

is the kth partial Fourier sum of Δ. In order to simplify our further calculations we shift Δ by π and obtain

$$\varphi(\alpha) := \Delta(\alpha+\pi) = \left|\sin\frac{\alpha}{2}\right| \Big/ \sqrt{1+\sin^2\frac{\alpha}{2}}\,.$$

The function φ is even and, therefore, its Fourier coefficients associated with $\sin\nu\alpha$ are all equal to zero. Let

$$a_0 := \frac{1}{2\pi}\int_0^{2\pi}\varphi(\alpha)\,d\alpha \quad \text{and} \quad a_\nu := \frac{1}{\pi}\int_0^{2\pi}\varphi(\alpha)\cos\nu\alpha\,d\alpha, \quad \nu = 1, 2, \dots$$

be the Fourier coefficients of φ associated with $\cos\nu\alpha$. Obviously, $a_0 > 0$. Let $\nu = 1, 2, \dots$. Then using integration by parts (twice) we get

$$\begin{aligned} a_\nu &= -\frac{1}{\pi\nu}\int_0^{2\pi}\varphi'(\alpha)\sin\nu\alpha\,d\alpha \\ &= \frac{1}{\pi\nu^2}\left[\varphi'(2\pi^-) - \varphi'(0^+)\right] - \frac{1}{\pi\nu^2}\int_0^{2\pi}\varphi''(\alpha)\cos\nu\alpha\,d\alpha \\ &= \frac{1}{\pi\nu^2}\int_0^{2\pi}\varphi''(\alpha)(1-\cos\nu\alpha)\,d\alpha. \end{aligned}$$

Therefore,

$$a_\nu = \frac{2}{\pi\nu^2}\int_0^{2\pi}\varphi''(\alpha)\sin^2\frac{\nu\alpha}{2}\,d\alpha. \tag{4.16}$$

Simple calculations show that

$$\varphi'(\alpha) = \frac{1}{2}\cos\frac{\alpha}{2}\Big/\left(1+\sin\frac{\alpha}{2}\right)^{3/2} \text{ and } \varphi''(\alpha) < 0 \text{ for } 0 < \alpha < 2\pi.$$

This and (4.16) imply that $a_\nu < 0$ and

$$|a_\nu| \le \frac{2}{\pi\nu^2}\int_0^{2\pi}|\varphi''(\alpha)|\,d\alpha = -\frac{2}{\pi\nu^2}\int_0^{2\pi}\varphi''(\alpha)\,d\alpha = -\frac{2}{\pi\nu^2}\left[\varphi'(2\pi^-) - \varphi'(0^+)\right] = \frac{2}{\pi\nu^2}.$$

Thus, we have

$$-\frac{2}{\pi\nu^2} \le a_\nu < 0, \quad \nu = 1, 2, \dots. \tag{4.17}$$

Therefore $\varphi(\alpha) = a_0 + \sum_{\nu=1}^{\infty} a_\nu \cos \nu\alpha$, where $a_0 > 0$ and $a_\nu < 0$, $\nu = 1, 2, \dots$, and hence

$$S_k(\varphi)(\alpha) = a_0 + \sum_{\nu=1}^{k} a_\nu \cos \nu\alpha \geq a_0 + \sum_{\nu=1}^{k} a_\nu = S_k(\varphi)(0)$$

$$= S_k(\varphi)(0) - \varphi(0) = -\sum_{\nu=k+1}^{\infty} a_\nu > 0.$$

Thus $S_k(\varphi)(\alpha) > 0$ for $\alpha \in [-\pi, \pi)$ and hence $S_k(\Delta)(\alpha) > 0$ for $\alpha \in [-\pi, \pi)$ which implies the lower bound in (4.15).

The inequalities (4.17) imply

$$\|\Delta - S_k(\Delta)\|_C = \|\varphi - S_k(\varphi)\|_C \leq ck^{-1}.$$

Using this, we obtain

$$\begin{aligned} |\eta_j| &\leq ck^{-1}|S_k(\Delta)(\alpha_j)| \leq ck^{-1}\left(|\Delta(\alpha_j)| + \|\Delta - S_k(\Delta)\|_C\right) \\ &\leq ck^{-1}\left(\cos\frac{\alpha_j}{2} + k^{-1}\right) \leq ck^{-1}\left(\cos\frac{\alpha_j}{2} + \cos\frac{\alpha_k}{2}\right) \leq ck^{-1}\cos\frac{\alpha_j}{2}, \end{aligned}$$

where we used that $\cos(\alpha_k/2) = \cos(\pi k/(2k+1)) > ck^{-1}$. Thus the upper estimate in (4.15) is proved. Lemma 4.1 is proved. □

REMARK 4.1. *The exact values of the nodes β_j and the coefficients λ_j of the quadrature* (4.5) *from Lemma* 4.1 *are the following:*

$$\beta_j = \frac{\pi}{2} - \arccos\left(\cos^2\frac{(k-j)\pi}{2k+1}\right), \quad j = 0, 1, \dots, k,$$

and $\beta_j = \pi - \beta_{2k-j}$, $\quad j = k+1, k+2, \dots, 2k$;

$$\lambda_j = \frac{2}{2k+1}\int_{-\pi}^{\pi} \cos\frac{\alpha}{2}\left(1 + \cos^2\frac{\alpha}{2}\right)^{-1/2} D_k\left(\alpha - \frac{2\pi(k-j)}{2k+1}\right) d\alpha, \quad j = 0, 1, \dots, k,$$

and $\lambda_j = \lambda_{2k-j}$, $\quad j = k+1, k+2, \dots, 2k$, *where D_k is the Dirichlet kernel of degree k.*

We are now in a position to construct our cubature formula for integration over $\mathbf{S}^{d-1}$ ($d > 2$). We shall use (4.2), (4.3), and the quadratures from (4.4) and (4.5).

Definition of cubature $\mathbf{Q}_n$. Given $n = 1, 2, \dots$ we select $k := 2(d-1)n + d - 2$. Let $\mathcal{J}_n$ be the set of all indices $\mathbf{j} := (j_1, \dots, j_{d-1})$ such that $0 \leq j_\nu \leq 2k$; i.e., $\mathcal{J}_n := \{0, 1, \dots, 2k\}^{d-1}$. Note that the cardinality of $\mathcal{J}_n$ is $\#\mathcal{J}_n = (2k+1)^{d-1} \asymp n^{d-1}$. Set $\beta_{\mathbf{j}} := (\beta_{j_1}, \dots, \beta_{j_{d-2}})$, $\gamma_{\mathbf{j}} := \gamma_{j_{d-1}}$, $\omega_{\mathbf{j}} := \xi(\beta_{\mathbf{j}}, \gamma_{\mathbf{j}})$, and $\lambda_{\mathbf{j}} := J(\beta_{\mathbf{j}})\varrho_{j_{d-1}}\prod_{\nu=1}^{d-2}\lambda_{j_\nu}$, where γ_j, ϱ_j, β_j, and λ_j are the nodes and the coefficients of quadratures (4.4) and (4.5), respectively, and J is from (4.1). We define

$$\mathbf{Q}_n(f) := \sum_{\mathbf{j}\in\mathcal{J}_n} \lambda_{\mathbf{j}} f(\omega_{\mathbf{j}}) \quad \sim \quad \int_{\mathbf{S}^{d-1}} f(\xi)\, d\xi. \tag{4.18}$$

When it is possible we shall write this cubature with the following simpler indices. Let Ω_n be the set of all nodes $\omega = \omega_{\mathbf{j}}$, and $\lambda_\omega := \lambda_{\omega_{\mathbf{j}}} := \lambda_{\mathbf{j}}$, $\mathbf{j} \in \mathcal{J}_n$. Then cubature (4.18) can be rewritten in the form

$$\mathbf{Q}_n(f) := \sum_{\omega\in\Omega_n} \lambda_\omega f(\omega) \quad \sim \quad \int_{\mathbf{S}^{d-1}} f(\xi)\, d\xi. \tag{4.19}$$

Observe that $\#\Omega_n \asymp n^{d-1}$.

As we mentioned in the beginning of this section, every spherical polynomial of degree $2n$ can be represented in spherical coordinates as a linear combination of terms like those in (4.3) and the Jacobian J is represented in a similar way (see (4.1)). On the other hand, quadratures (4.4) and (4.5) are exact for trigonometric polynomials of degree $k := 2(d-1)n + d - 2$. Therefore (see (4.2)), cubature (4.18) (or (4.19)) is exact for all spherical polynomials of degree $2n$; i.e., for every spherical polynomial S of degree $\le 2n$ we have

$$\mathbf{Q}_n(S) := \sum_{\omega\in\Omega_n} \lambda_\omega S(\omega) = \int_{\mathbf{S}^{d-1}} S(\xi)\,d\xi. \tag{4.20}$$

Note that $\lambda_\omega > 0$ and, since (4.20) holds for $S = 1$, then

$$\sum_{\omega\in\Omega_n} \lambda_\omega = \int_{\mathbf{S}^{d-1}} 1\,d\xi =: |\mathbf{S}^{d-1}|. \tag{4.21}$$

Identity (4.20) implies discrete representations of the projection $Q_m(f)$ of any $f \in L_2(\mathbf{B}^d)$ onto $\mathcal{P}_m \ominus \mathcal{P}_{m-1}$ and $\|A_m(f)\|_{L_2(\mathbf{S}^{d-1})}$ (see (3.6) and (3.7) from Theorem 3.1). Namely, since $A_m(f,\xi)\mathcal{U}_m(\mathbf{x}\cdot\xi)$ and $A_m^2(f,\xi)$, for $m \le n$, are spherical polynomials of degree $\le 2m \le 2n$, then

$$Q_m(f,\mathbf{x}) := \nu_m \int_{\mathbf{S}^{d-1}} A_m(f,\xi)\mathcal{U}_m(\mathbf{x}\cdot\xi)\,d\xi = \nu_m \sum_{\omega\in\Omega_n} \lambda_\omega A_m(f,\omega)\mathcal{U}_m(\mathbf{x}\cdot\omega) \tag{4.22}$$

and

$$\|A_m\|^2_{L_2(\mathbf{S}^{d-1})} := \int_{\mathbf{S}^{d-1}} |A_m(f,\xi)|^2\,d\xi = \sum_{\omega\in\Omega_n} \lambda_\omega |A_m(f,\omega)|^2. \tag{4.23}$$

Since quadratures (4.4) and (4.5) have good localization properties, then cubature (4.18) (or (4.19)) has such properties. We shall use them to prove the following lemma.

LEMMA 4.2. *Let $n = 1, 2, \dots$, $m = 1, 2, \dots$, and let, for $\gamma \in (0, \pi]$, $\mathcal{K}_m(\cos\gamma) := c_0 \min\{m^{d-1}, m^{d-1}/(m\gamma)^d\}$ with $c_0 > 0$ a constant. Then we have*

$$\mathbf{Q}_n(\mathcal{K}_m(\bullet\cdot\eta)) \le c[1 + (m/n)^{d-1}] \quad \text{for } \eta \in \mathbf{S}^{d-1}, \tag{4.24}$$

where $\mathbf{Q}_n$ is the cubature from (4.19) and c depends only on d and c_0.

Proof. In what follows, we shall assume that $n > n_0$, where n_0 is sufficiently large and depends only on the dimension d. Estimate (4.24) obviously holds for $n \le n_0$ by (4.21). We first construct a tiling of $\mathbf{S}^{d-1}$ which is determined by the nodes of cubature (4.18). We associate with each node $\omega_{\mathbf{j}}$ the spherical box (tile) $T_{\mathbf{j}}$ consisting of all points $\xi \in \mathbf{S}^{d-1}$ for which $\xi = \xi(\theta,\phi)$ with

$$(\theta,\phi) \in [a_{j_1}, a_{j_1+1}) \times \cdots \times [a_{j_{d-2}}, a_{j_{d-2}+1}) \times [b_{j_{d-1}}, b_{j_{d-1}+1}),$$

where $a_j := \frac{1}{2}(\beta_j + \beta_{j-1})$ and $b_j := \frac{1}{2}(\gamma_j + \gamma_{j-1})$ with β_j from (4.5) and γ_j from (4.4). Observe that $\omega_{\mathbf{j}} \in T_{\mathbf{j}}$ is the (spherical) center of $T_{\mathbf{j}}$. Obviously $T_{\mathbf{j}} \bigcap T_{\mathbf{i}} = \emptyset$, $\mathbf{j} \ne \mathbf{i}$, and the tiles $T_{\mathbf{j}}$ cover $\mathbf{S}^{d-1}$ excluding small regions around the poles. The most important property of our cubature is that

$$0 < \lambda_{\mathbf{j}} \le c \int_{T_{\mathbf{j}}} 1\,d\xi =: c|T_{\mathbf{j}}| \quad \text{for } \mathbf{j} \in \mathcal{J}_n. \tag{4.25}$$

This property follows readily by (4.7), the definition of γ_j from (4.4), and the definition of our cubature (see (4.18)).

The second important property of our tiling is that the diameter of each tile $T_{\mathbf{j}}$ is $\le cn^{-1}$. We let $\rho(\xi,\eta) := \arccos \xi\cdot\eta,\ \xi,\eta\in\mathbf{S}^{d-1}$ denote the angular distance on $\mathbf{S}^{d-1}$ (the angle between vectors ξ and η). It is easily seen that $\rho(\xi,\eta)$ satisfies the axioms for a distance on $\mathbf{S}^{d-1}$. Since $\beta_j-\beta_{j-1}\le cn^{-1}$, by (4.6), and $\gamma_j-\gamma_{j-1}\le cn^{-1}$, by the definition of γ_j, then

$$\sup\{\rho(\xi,\eta):\xi,\eta\in T_{\mathbf{j}}\} < c_1 n^{-1}, \tag{4.26}$$

where c_1 depends only on d.

Suppose that $\eta\in\mathbf{S}^{d-1}$ is fixed. We select a new coordinate system such that $\eta = \mathbf{e}_1' := (1,0,\dots,0)$ is its first coordinate vector. This can be done by a suitable rotation of the old coordinate system. For $\xi\in\mathbf{S}^{d-1}$, we shall denote by $\theta' := (\theta_1',\dots,\theta_{d-2}')$ and ϕ' the new spherical coordinates of ξ.

We define, for $\nu = 1,2,\dots,n$,

$$\begin{aligned}\mathcal{Z}_\nu &:= \left\{\xi\in\mathbf{S}^{d-1} : \frac{\pi(\nu-1)}{n}\le \rho(\xi,\mathbf{e}_1')\le\frac{\pi\nu}{n}\right\}\\ &= \left\{\xi\in\mathbf{S}^{d-1} : \frac{\pi(\nu-1)}{n}\le \theta_1'\le\frac{\pi\nu}{n}\right\}\end{aligned}$$

and

$$\mathcal{Z}_\nu^* := \left\{\xi\in\mathbf{S}^{d-1} : \max\left\{\frac{\pi(\nu-1)-c_1}{n},\,-\pi\right\}\le\theta_1'\le\min\left\{\frac{\pi\nu+c_1}{n},\,\pi\right\}\right\},$$

where c_1 is from (4.26). Obviously $\bigcup_{\nu=1}^n \mathcal{Z}_\nu = \mathbf{S}^{d-1}$.

Let $\mathcal{T}_\nu$ be the set of all tiles $T_{\mathbf{j}}$ with centers $\omega_{\mathbf{j}}\in\mathcal{Z}_\nu$. It follows by (4.26) that $\bigcup_{T\in\mathcal{T}_\nu} T\subset\mathcal{Z}_\nu^*$ and hence

$$\sum_{T\in\mathcal{T}_\nu}|T| \le |\mathcal{Z}_\nu^*| := \int_{\mathcal{Z}_\nu^*} 1\,d\xi \le c\int_{\mathcal{Z}_\nu} 1\,d\xi =: c|\mathcal{Z}_\nu|,\quad \nu=1,2,\dots,n. \tag{4.27}$$

We are now ready to estimate $\mathbf{Q}_n(\mathcal{K}_m(\bullet\cdot\eta))$. If $\nu=1$, then we obtain, using (4.25), (4.27), and the assumptions of the lemma,

$$\begin{aligned}&\sum_{\omega_{\mathbf{j}}\in\mathcal{Z}_1}\lambda_{\mathbf{j}}\mathcal{K}_m(\omega_{\mathbf{j}}\cdot\mathbf{e}_1') \le c\max\{\mathcal{K}_m(\cos\theta_1):0\le\theta_1'\le\pi/n\}\sum_{T\in\mathcal{T}_1}|T|\\ &\le cm^{d-1}|\mathcal{Z}_1| \le cm^{d-1}\int_0^{\pi/n}\sin^{d-2}\theta_1'\,d\theta_1' \le c(m/n)^{d-1}.\end{aligned}$$

If $\nu\ge 2$, then

$$\begin{aligned}&\sum_{\omega_{\mathbf{j}}\in\mathcal{Z}_\nu}\lambda_{\mathbf{j}}\mathcal{K}_m(\omega_{\mathbf{j}}\cdot\mathbf{e}_1') \le c\max\{\mathcal{K}_m(\cos\theta_1'):\pi(\nu-1)/n\le\theta_1'\le\pi\nu/n\}\sum_{T\in\mathcal{T}_\nu}|T|\\ &\le c\mathcal{K}_m\left(\cos\frac{\pi(\nu-1)}{n}\right)|\mathcal{Z}_\nu^*| \le c\mathcal{K}_m\left(\cos\frac{\pi\nu}{n}\right)|\mathcal{Z}_\nu| \le c\int_{\mathcal{Z}_\nu}\mathcal{K}_m(\xi\cdot\mathbf{e}_1')\,d\xi,\end{aligned}$$

where we used that $\mathcal{K}_m(\cos\frac{\pi(\nu-1)}{n}) \le c\mathcal{K}_m\left(\cos\frac{\pi\nu}{n}\right)$, $\nu\ge 2$, which follows by the definition of $\mathcal{K}_m(\cos\gamma)$ from the assumptions of the lemma. Therefore,

$$\sum_{\mathbf{j}\in\mathcal{J}_n}\lambda_{\mathbf{j}}\mathcal{K}_m(\omega_{\mathbf{j}}\cdot\mathbf{e}_1') \le c(m/n)^{d-1} + c\int_{\mathcal{Z}}\mathcal{K}_m(\xi\cdot\mathbf{e}_1')\,d\xi, \tag{4.28}$$

where $\mathcal{Z} := \bigcup_{\nu=2}^{n} \mathcal{Z}_\nu$. We obtain, using again the definition of $\mathcal{K}_m(\cos\gamma)$,

$$\int_{\mathcal{Z}} \mathcal{K}_m(\xi\cdot\mathbf{e}_1')\,d\xi = |\mathbf{S}^{d-2}| \int_{\pi/n}^{\pi} \mathcal{K}_m(\cos\theta_1')\sin^{d-2}\theta_1'\,d\theta_1'$$
$$\le c\int_0^{\infty} \min\{m^{d-1}, m^{d-1}/(m\theta_1')^d\}(\theta_1')^{d-2}\,d\theta_1' \le c < \infty.$$

The above estimates and (4.28) imply (4.24). □

We shall deal with discrete sums of spherical polynomial values. For this, we need a rapidly decaying reproducing kernel for the space of spherical polynomials of degree m. The following well-known proposition gives us such a kernel.

PROPOSITION 4.1. *There exists a constant $m_0 = m_0(d)$ such that for every $m \ge m_0$ there exists an algebraic polynomial W_m of degree dm with the properties:*

(a)

$$S(\eta) = \int_{\mathbf{S}^{d-1}} W_m(\eta\cdot\xi)S(\xi)\,d\xi, \quad \eta\in\mathbf{S}^{d-1},$$

for each spherical polynomial S of degree $\le m$;

(b)

$$|W_m(\cos\tau)| \le c_0 \min\{m^{d-1}, m^{d-1}/(m\tau)^d\} \quad \textit{for} \quad 0<\tau\le\pi, \tag{4.29}$$

and hence

(c)

$$\int_{\mathbf{S}^{d-1}} |W_m(\eta\cdot\xi)|\,d\xi \le c < \infty, \quad \eta\in\mathbf{S}^{d-1}, \tag{4.30}$$

where c_0 and c are independent of m and η.

Since we do not have a good reference for Proposition 4.1, we shall show how it can be deduced from the following results of Kogbetliantz and Stein (see also [P]).

PROPOSITION 4.2 (see [K]). *Let $S_m(t) := \sum_{\nu=0}^{m}(\nu+\lambda)C_\nu^\lambda(t)$, $\lambda > 0$, $m = 0,1,\dots,$ and let $\sigma_m^{(\delta)}$ be the Cesàro means of order δ of S_m; i.e.,*

$$\sigma_m^{(\delta)}(t) := \left(A_m^\delta\right)^{-1}\sum_{\nu=0}^{m} A_{m-\nu}^\delta(\nu+\lambda)C_\nu^\lambda(t) \quad \textit{with } A_\nu^\delta := \tfrac{\Gamma(\nu+\delta+1)}{\Gamma(\delta+1)\Gamma(\nu+1)}. \tag{4.31}$$

Then, for $-1<\delta\le 2\lambda+1$,

$$|\sigma_m^{(\delta)}(\cos\gamma)| \le c\min\left\{(m+1)^{2\lambda+1}, (m+1)^{2\lambda-\delta}\Big/\left(\sin\frac{\gamma}{2}\right)^{\delta+1}\right\},\ 0<\gamma\le\pi, \tag{4.32}$$

with c depending only on λ.

PROPOSITION 4.3 (see [St]). *For each positive integer r and for $m = 0,1,\dots,$ there exist $r+1$ parameters $\alpha_1(m),\dots,\alpha_{r+1}(m)$ (depending only on m and r) which are uniformly bounded: $|\alpha_\nu(m)|\le A$, A independent of m, and there exists a fixed integer N, so that the following holds:*

If $\sum_{\nu=0}^{\infty} a_\nu$ is a series of real numbers and if $\sigma_m^{(r)}$, $m=0,1,\dots,$ are the Cesàro means of order r of the partial sums S_m, $m=0,1,\dots,$ of this series (see (4.31)), then

$$\tau_m^{(r)} := \alpha_1(m)\sigma_{m-1}^{(r)} + \alpha_2(m)\sigma_{2m-1}^{(r)} + \cdots + \alpha_{r+1}(m)\sigma_{(r+1)m-1}^{(r)}$$

can be represented in the form

$$\tau_m^{(r)} = \sum_{\nu=0}^{m} a_\nu + \sum_{\nu=m+1}^{(r+1)m} \beta_\nu a_\nu \quad \text{if } m \ge N,$$

where β_ν are constants depending on m and r.

Proof of Proposition 4.1. We have already mentioned in (3.15) that

$$K_m(t) := \frac{N(d,m)}{|\mathbf{S}^{d-1}| C_m^{(d-2)/2}(1)} C_m^{(d-2)/2}(t)$$

gives the reproducing kernel $K_m(\xi \cdot \eta)$ for $\mathcal{H}_m$. Therefore, $\sum_{\nu=0}^m K_\nu(\xi \cdot \eta)$ is a reproducing kernel for all spherical polynomials of degree $\le m$. Simple calculations show that

$$K_m(t) = 2\left[|\mathbf{S}^{d-1}|(d-2)\right]^{-1}(m+\lambda)C_m^\lambda(t) \quad \text{with } \lambda := (d-2)/2.$$

Therefore, $2\left[|\mathbf{S}^{d-1}|(d-2)\right]^{-1}\sum_{\nu=0}^m(\nu+\lambda)C_\nu^\lambda(t)$ gives a reproducing kernel for the spherical polynomials of degree $\le m$.

We now apply Proposition 4.2 with $\lambda := (d-2)/2$ and $\delta := 2\lambda + 1 = d-1$. Then we apply Proposition 4.3 to the resulting Cesàro means $\{\sigma_\nu^{(r)}\}$ with $r := \delta = d-1$ to conclude that $W_m := 2\left[|\mathbf{S}^{d-1}|(d-2)\right]^{-1}\tau_m^{(r)}$ satisfies (4.29) (by (4.32) and since $\alpha_\nu(m)$ are uniformly bounded) and $W_m(\xi \cdot \eta)$ is a reproducing kernel for the spherical polynomials of degree $\le m$ (by Proposition 4.3). □

Lemma 4.2 and Proposition 4.1 allow us to estimate discrete $l_p(\Omega_n)$ norms of spherical polynomials by their $L_p(\mathbf{S}^{d-1})$ norms. In this part we use ideas from [O].

LEMMA 4.3. *Let $n = 1, 2, \dots$, and let $m \ge m_0$, where m_0 is from Proposition* 4.1. *Then for every spherical polynomial S of degree m and for $1 \le p \le \infty$ we have*

$$\sum_{\omega \in \Omega_n} \lambda_\omega |S(\omega)|^p \le c[1 + (m/n)^{d-1}] \int_{\mathbf{S}^{d-1}} |S(\xi)|^p \, d\xi, \tag{4.33}$$

where λ_ω and Ω_n are from (4.19), *and c is independent of S, n, and m.*

Proof. By Proposition 4.1 we get $S(\omega) = \int_{\mathbf{S}^{d-1}} W_m(\omega \cdot \xi) S(\xi)\, d\xi$, $\omega \in \Omega_n$. We obtain, using Hölder's inequality,

$$\begin{aligned} |S(\omega)| &\le \int_{\mathbf{S}^{d-1}} |W_m(\omega \cdot \xi) S(\xi)|\, d\xi = \int_{\mathbf{S}^{d-1}} |W_m(\eta \cdot \xi)|^{1-1/p} |W_m(\omega \cdot \xi)|^{1/p} |S(\xi)|\, d\xi \\ &\le \left(\int_{\mathbf{S}^{d-1}} |W_m(\omega \cdot \xi)|\, d\xi \right)^{1-1/p} \left(\int_{\mathbf{S}^{d-1}} |W_m(\omega \cdot \xi)| |S(\xi)|^p \, d\xi \right)^{1/p} \end{aligned}$$

and hence

$$|S(\omega)|^p \le A^{p-1} \int_{\mathbf{S}^{d-1}} |W_m(\omega \cdot \xi)| |S(\xi)|^p \, d\xi, \quad \text{where } A := \textstyle\int_{\mathbf{S}^{d-1}} |W_m(\omega \cdot \xi)|\, d\xi.$$

We now multiply both sides of the above inequality by λ_ω and sum over $\omega \in \Omega_n$ to obtain

$$\begin{aligned} &\sum_{\omega \in \Omega_n} \lambda_\omega |S(\omega)|^p \le A^{p-1} \int_{\mathbf{S}^{d-1}} \left(\sum_{\omega \in \Omega_n} \lambda_\omega |W_m(\omega \cdot \xi)| \right) |S(\xi)|^p \, d\xi \\ &\le A^{p-1} \max_{\xi \in \mathbf{S}^{d-1}} \mathbf{Q}_n(|W_m(\bullet \cdot \xi)|) \int_{\mathbf{S}^{d-1}} |S(\xi)|^p \, d\xi. \end{aligned}$$

It follows, by (4.29), that $|W_m(\bullet\cdot\xi)| \le \mathcal{K}_m(\bullet\cdot\xi)$, where $\mathcal{K}_m$ is defined in Lemma 4.2 with c_0 from Proposition 4.1. Then Proposition 4.1 and Lemma 4.2 imply

$$\max_{\xi\in\mathbf{S}^{d-1}} \mathbf{Q}_n(|W_m(\bullet\cdot\xi)|) \le \max_{\xi\in\mathbf{S}^{d-1}} \mathbf{Q}_n(\mathcal{K}_m(\bullet\cdot\xi)) \le c[1+(m/n)^{d-1}] \quad \text{and} \quad A \le c$$

which completes the proof of Lemma 4.3. □

The following lemma relates the $L_2(\mathbf{S}^{d-1})$ norms and discrete $l_2(\Omega_n)$ norms of spherical polynomials written in terms of $\nu_m\mathcal{U}_m(\xi\cdot\omega)/\mathcal{U}_m(1)$, the reproducing kernel for the space $\mathcal{H}_m \oplus \mathcal{H}_{m-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$ (see (3.17)).

LEMMA 4.4. *Let $n = 1, 2, \ldots$, and let $c(\omega)$, $\omega \in \Omega_n$, be real constants. Let $m \ge m_0$, where m_0 is from Proposition* 4.1. *Then the spherical polynomial*

$$S(\xi) := \sum_{\omega\in\Omega_n} \lambda_\omega c(\omega) \frac{\nu_m}{\mathcal{U}_m(1)} \mathcal{U}_m(\xi\cdot\omega)$$

satisfies

$$\|S\|^2_{L_2(\mathbf{S}^{d-1})} \le c[1+(m/n)^{d-1}] \sum_{\omega\in\Omega_n} \lambda_\omega |c(\omega)|^2. \tag{4.34}$$

Proof. Using (3.11) we get

$$\begin{aligned}
\|S\|^2_{L_2(\mathbf{S}^{d-1})} &= \int_{\mathbf{S}^{d-1}} |S(\xi)|^2\, d\xi \\
&= \sum_{\omega\in\Omega_n} \sum_{\eta\in\Omega_n} \lambda_\omega\lambda_\eta c(\omega)c(\eta) \left(\frac{\nu_m}{\mathcal{U}_m(1)}\right)^2 \int_{\mathbf{S}^{d-1}} \mathcal{U}_m(\xi\cdot\omega)\mathcal{U}_m(\xi\cdot\eta)\, d\xi \\
&= \sum_{\omega\in\Omega_n} \sum_{\eta\in\Omega_n} \lambda_\omega\lambda_\eta c(\omega)c(\eta) \frac{\nu_m}{\mathcal{U}_m(1)} \mathcal{U}_m(\omega\cdot\eta) = \sum_{\eta\in\Omega_n} \lambda_\eta c(\eta) S(\eta) \\
&\le \left(\sum_{\eta\in\Omega_n} \lambda_\eta |c(\eta)|^2\right)^{1/2} \left(\sum_{\eta\in\Omega_n} \lambda_\eta |S(\eta)|^2\right)^{1/2}.
\end{aligned}$$

By Lemma 4.3, the last quantity above does not exceed $c[1+(m/n)^{d-1}]^{\frac12} \|S\|_{L_2(\mathbf{S}^{d-1})}$. Finally, we divide by $\|S\|_{L_2(\mathbf{S}^{d-1})}$ to complete the proof of the lemma. □

5. Smoothness spaces in $L_2(\mathbf{B}^d)$. In this section, we shall recall results about approximation by algebraic polynomials. As earlier, we let $\mathcal{P}_n$ denote the space of algebraic polynomials in d-variables. For $n \ge 1$, let

$$E_n(f) := E_n(f)_{L_2(\mathbf{B}^d)} := \inf_{P\in\mathcal{P}_n} \|f - P\|_{L_2(\mathbf{B}^d)}$$

be the error in approximating $f \in L_2(\mathbf{B}^d)$ by algebraic polynomials P of degree $\le n$. By Theorem 3.1 we have the following representation of the polynomial $P_n(f,\mathbf{x})$ of best $L_2(\mathbf{B}^d)$-approximation to f:

$$P_n(f,\mathbf{x}) = \sum_{m=0}^{n} \nu_m \int_{\mathbf{S}^{d-1}} A_m(\xi)\mathcal{U}_m(\mathbf{x}\cdot\xi)\, d\xi, \tag{5.1}$$

where

$$A_n(\xi) := A_n(f,\xi) := \int_{\mathbf{B}^d} f(\mathbf{y})\mathcal{U}_m(\mathbf{y}\cdot\xi)\, d\mathbf{y}.$$

Since $A_m(\xi)U_m(\mathbf{x}\cdot\xi)$ is a spherical polynomial of degree $\le 2m \le 2n$ in ξ, we can use the quadrature formula (4.19) to obtain

$$P_n(f,\mathbf{x}) = \sum_{\omega\in\Omega_n} \lambda_\omega \sum_{m=0}^{n} \nu_m A_m(\omega)\mathcal{U}_m(\mathbf{x}\cdot\omega). \tag{5.2}$$

From Theorem 3.1, we have

$$\begin{aligned} E_n(f)^2 &= \|f - P_n(f)\|^2_{L_2(\mathbf{B}^d)} = \sum_{m>n} \nu_m \|A_m(f)\|^2_{L_2(\mathbf{S}^{d-1})} \\ &\asymp \sum_{m>n} m^{d-1}\|A_m(f)\|^2_{L_2(\mathbf{S}^{d-1})}. \end{aligned} \tag{5.3}$$

For $\alpha > 0$, let $W^\alpha(L_2(\mathbf{B}^d))$ be the Sobolev space for the domain $\mathbf{B}^d$. When $\alpha = k$ is an integer, then a function $f \in L_2(\mathbf{B}^d)$ is in $W^k(L_2(\mathbf{B}^d))$ if and only if its distributional derivatives $D^\nu f$ of order k are in $L_2(\mathbf{B}^d)$, and

$$|f|^2_{W^k(L_2(\mathbf{B}^d))} := \sum_{|\nu|=k} \|D^\nu f\|^2_{L_2(\mathbf{B}^d)}$$

gives the seminorm for $W^k(L_2(\mathbf{B}^d))$. The norm for $W^k(L_2(\mathbf{B}^d))$ is obtained by adding $\|f\|_{L_2(\mathbf{B}^d)}$ to $|f|_{W^k(L_2(\mathbf{B}^d))}$. For other values of α, we obtain W^α as the interpolation space

$$W^\alpha(L_2(\mathbf{B}^d)) = (L_2(\mathbf{B}^d), W^k(L_2(\mathbf{B}^d)))_{\theta,2}, \quad \theta = \alpha/k,\ 0 < \alpha < k,$$

given by the real method of interpolation (see, e.g., Bennett and Sharpley [BS]).

A fundamental result in approximation known as the Jackson theorem states that

$$E_n(f) \le c(k)n^{-k}\|f\|_{W^k(L_2(\mathbf{B}^d))}, \tag{5.4}$$

where the norm on the right can be replaced by the seminorm if k is an integer. This theorem can be deduced easily from the results on univariate approximation in Chapter 7 of [DL]. By interpolation (see, e.g., [DL, Chapter 7]), one obtains

$$\sum_{n=1}^{\infty} [n^\alpha E_n(f)]^2 n^{-1} \le c(\alpha)\|f\|^2_{W^\alpha(L_2(\mathbf{B}^d))}, \quad \alpha > 0, \tag{5.5}$$

with $c(\alpha)$ depending at most on α. From (5.3) and (5.5), it is easy to deduce that

$$\sum_{n=1}^{\infty} n^{2\alpha+d-1}\|A_n(f)\|^2_{L_2(\mathbf{S}^{d-1})} \le c(\alpha)\|f\|^2_{W^\alpha(L_2(\mathbf{B}^d))}, \quad \alpha > 0, \tag{5.6}$$

with $c(\alpha)$ depending at most on α.

6. Approximation of functions in $L_2(I, w)$. We shall also need certain results about the approximation of univariate functions in $L_2(I,w)$ where $I := [-1,1]$ and $w := w_{d/2}$. As we know by section 2, the Gegenbauer polynomials $\{\mathcal{U}_m\}_{m=0}^\infty$ form a complete orthonormal system for $L_2(I,w)$ (see (3.3)). For any $g \in L_2(I,w)$ we have

$$g = \sum_{m=0}^{n} \hat{g}(m)\mathcal{U}_m \quad \text{with} \quad \hat{g}(m) := \int_I g(s)\mathcal{U}_m(s)w(s)\,ds. \tag{6.1}$$

We shall use approximation of functions in $L_2(I,w)$ as an intermediate tool in establishing our results on ridge approximation. Let $\mathcal{P}_n(I)$ denote the space of univariate algebraic polynomials of degree $\le n$. For a function $g \in L_2(I,w)$, we let

$$E_n(g)_{L_2(I,w)} := \inf_{p\in\mathcal{P}_n(I)} \|g-p\|_{L_2(I,w)}$$

be the error in approximating g by the elements of $\mathcal{P}_n(I)$. The polynomial

$$p_n := \sum_{m=0}^{n} \hat{g}(m)\mathcal{U}_m \tag{6.2}$$

is the best $L_2(I,w)$ approximation to g by elements of $\mathcal{P}_n(I)$, and we have

$$E_n(g)^2_{L_2(I,w)} = \|g-p_n\|^2_{L_2(I,w)} = \sum_{m>n} |\hat{g}(m)|^2. \tag{6.3}$$

We introduce the univariate Sobolev spaces $W^\alpha(L_2(I,w))$, $\alpha \in \mathbf{R}$, whose norms are defined by

$$\|g\|^2_{W^\alpha(L_2(I,w))} := \sum_{m=0}^{\infty} [(m+1)^\alpha |\hat{g}(m)|]^2. \tag{6.4}$$

It follows that for each $g \in W^\alpha(L_2(I,w))$,

$$E_n(g)_{L_2(I,w)} \le c(\alpha) n^{-\alpha} \|g\|_{W^\alpha(L_2(I,w))}. \tag{6.5}$$

Moreover, similar to (5.5), we have

$$\sum_{n=1}^{\infty} [n^\alpha E_n(g)_{L_2(I,w)}]^2 n^{-1} \le c(\alpha)\|g\|^2_{W^\alpha(L_2(I,w))}, \quad \alpha > 0. \tag{6.6}$$

There is also a Bernstein-type inequality for polynomials in $\mathcal{P}_n(I)$ with respect to $L_2(I,w)$ which follows trivially from the definition: for every $p \in \mathcal{P}_n(I)$ and $\alpha > 0$,

$$\|p\|_{W^\alpha(L_2(I,w))} \le (n+1)^\alpha \|p\|_{L_2(I,w)}. \tag{6.7}$$

It is well known (see [DL, Chapter 7]) that companion inequalities like (6.5) and (6.7) imply a characterization of approximation spaces by interpolation spaces. In our context, the approximation spaces are the Sobolev spaces $W^\alpha(L_2(I,w))$ defined by (6.4) and we therefore obtain for each $0 < \alpha < k$,

$$W^\alpha(L_2(I,w)) = (L_2(I,w), W^k(L_2(I,w)))_{\theta,2}, \quad \theta = \alpha/k. \tag{6.8}$$

Further properties of the spaces $W^\alpha(L_2(I,w))$ are given in section 7.

7. Approximation by ridge functions. In this section, we assume that X_n is a subspace of $L_2(I,w)$, $w = w_{d/2}$, of dimension n with the following property. There is a real number $s > 0$ such that, for each univariate function $g \in W^s(L_2(I,w))$, there is a function $r \in X_n$ which provides the Jackson estimate

$$\|g-r\|_{L_2(I,w)} \le c_0 n^{-s} \|g\|_{W^s(L_2(I,w))}, \tag{7.1}$$

with c_0 a constant independent of g and n.

We define Y_n to be the space of functions R in d variables of the form

$$R(\mathbf{x}) = \sum_{\omega\in\Omega_n} r_\omega(\mathbf{x}\cdot\omega), \quad r_\omega \in X_n, \quad \omega \in \Omega_n, \tag{7.2}$$

where Ω_n is the set of vectors in $\mathbf{S}^{d-1}$ from (4.19). Then Y_n is a linear space of dimension $\le n\#\Omega_n \le cn^d$. We prove the following theorem about approximation from Y_n.

THEOREM 7.1. *Let X_n, $n = 1, 2, \dots$, satisfy inequality (7.1) for some $s > 0$. If f is a function from the space $W^{s+\frac{d-1}{2}}(L_2(\mathbf{B}^d))$, then there is a function R in Y_n such that*

$$\|f - R\|_{L_2(\mathbf{B}^d)} \le cn^{-s-\frac{d-1}{2}} \|f\|_{W^{s+\frac{d-1}{2}}(L_2(\mathbf{B}^d))} \tag{7.3}$$

with c a constant depending only on s and d.

REMARK 7.1. *If $s + (d-1)/2 - 1$ is an integer and the space Y_n contains $\mathcal{P}_{s+(d-1)/2-1}$, then we have that $\|f\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))}$ can be replaced by the seminorm $|f|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))}$.*

An important element of the proof of Theorem 7.1 is the idea to get rid of the "low frequencies" when approximating. To this end we shall use the following geometric construction which was proven for us by Boris Kashin.

LEMMA 7.1. *Let H be a Hilbert space with norm $\|\cdot\|$ and let A, $B \subset H$ be finite-dimensional linear subspaces of H with $\dim A \le \dim B$. If there exists δ, $0 < \delta < 1/2$, such that*

$$\sup_{\substack{x\in A\\ \|x\|\le 1}} \inf_{y\in B} \|x - y\| \le \delta, \tag{7.4}$$

then there is a constant c depending only on δ and a linear operator $L : A \to B$ such that for every $x \in A$,

$$\|Lx - x\| \le c \inf_{y\in B} \|x - y\|, \tag{7.5}$$

and

$$Lx - x \perp A \quad (Lx - x \text{ is orthogonal to } A).$$

Proof. See [DOP, Lemma 6]. □

Proof of Theorem 7.1. Estimate (7.3) trivially holds if $n < m_0$, where $m_0 = m_0(d)$ is the constant from Proposition 4.1.

Suppose that $n \ge m_0$. Let $P = P_n$ be the polynomial in $\mathcal{P}_n$ given by (5.1) (or (5.2)). Since P is the best $L_2(\mathbf{B}^d)$ approximation of f, it satisfies (see (5.4))

$$\|f - P\|_{L_2(\mathbf{B}^d)} \le cn^{-s-(d-1)/2} \|f\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))} \tag{7.6}$$

with c and all subsequent constants in this proof depending only on s and d. We shall approximate P by an element R of Y_N, $N = k_0 n$, where k_0 is a sufficiently large constant depending only on s and d.

We have $A_m(P, \xi) = A_m(f, \xi)$, $m \le n$, and $A_m(P, \xi) = 0$, $m > n$. Since $f \in W^{s+(d-1)/2}(L_2(\mathbf{B}^d))$, we know from (5.6) that

$$\sum_{m=0}^{n} (m+1)^{2s+2(d-1)} \|A_m(f)\|^2_{L_2(\mathbf{S}^{d-1})} \le c\|f\|^2_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))}. \tag{7.7}$$

From this, using (4.23), we obtain

$$\sum_{m=0}^{n}(m+1)^{2s+2(d-1)}\sum_{\omega\in\Omega_n}\lambda_\omega|(A_m(f,\omega))|^2\le c\|f\|^2_{W^{r+(d-1)/2}(L_2(\mathbf{B}^d))}. \tag{7.8}$$

We introduce the univariate polynomials

$$p_\omega(t):=\sum_{m=0}^{n}\nu_m A_m(f,\omega)\mathcal{U}_m(t)=\sum_{m=0}^{n}\nu_m A_m(P,\omega)\mathcal{U}_m(t),\quad \omega\in\Omega_n. \tag{7.9}$$

We have, by (4.22) and (7.9),

$$P(\mathbf{x})=\sum_{m=0}^{n}\nu_m\sum_{\omega\in\Omega_n}\lambda_\omega A_m(P,\omega)\mathcal{U}_m(\mathbf{x}\cdot\omega)=\sum_{\omega\in\Omega_n}\lambda_\omega p_\omega(\mathbf{x}\cdot\omega). \tag{7.10}$$

According to (6.4), we have

$$\begin{aligned}\|\lambda_\omega^{1/2}p_\omega\|^2_{W^s(L_2(I,w))}&=\sum_{m=0}^{n}(m+1)^{2s}\nu_m^2\lambda_\omega|A_m(f,\omega)|^2\\&\asymp\sum_{m=0}^{n}(m+1)^{2s+2(d-1)}\lambda_\omega|A_m(f,\omega)|^2.\end{aligned}$$

Hence, from (7.8),

$$\begin{aligned}\sum_{\omega\in\Omega_n}\|\lambda_\omega^{1/2}p_\omega\|^2_{W^s(L_2(I,w))}&\le c\sum_{m=0}^{n}(m+1)^{2s+2(d-1)}\sum_{\omega\in\Omega_n}\lambda_\omega|(A_m(f,\omega))|^2\\&\le c\|f\|^2_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))}.\end{aligned} \tag{7.11}$$

We shall approximate each polynomial p_ω by elements of X_N. We apply Lemma 7.1 in the following setting. We take for H the Hilbert space $L_2(I,w)$ and take $A=\mathcal{P}_n(I)$ and $B=X_N$ with $N\ge k_0 n$ and k_0 a positive integer. We next show that if k_0 is large enough then the assumption (7.4) is satisfied. We mentioned earlier in (6.7) that $\mathcal{P}_n(I)$ satisfies the Bernstein inequality

$$\|p\|_{W^s(L_2(I,w))}\le(n+1)^s\|p\|_{L_2(I,w)},\quad p\in\mathcal{P}_n(I).$$

If $p\in\mathcal{P}_n(I)$, then from this Bernstein inequality and from (7.1), there is an $r\in X_N$ such that

$$\|p-r\|_{L_2(I,w)}\le c_0N^{-s}\|p\|_{W^s(L_2(I,w))}\le c_0N^{-s}2^sn^s\|p\|_{L_2(I,w)}\le c_0 2^sk_0^{-s}\|p\|_{L_2(I,w)}.$$

Thus, if k_0 is large enough, condition (7.4) is satisfied. Therefore, for each $\omega\in\Omega_n$, we can find $r_\omega\in X_N$ such that $r_\omega-p_\omega\perp\mathcal{P}_n(I)$ with respect to the inner product in $L_2(I,w)$ and, by (7.1) and (7.5),

$$\|p_\omega-r_\omega\|^2_{L_2(I,w)}\le cn^{-2s}\|p_\omega\|^2_{W^s(L_2(I,w))}.$$

Therefore,

$$r_\omega-p_\omega=\sum_{m=n+1}^{\infty}\hat{r}_\omega(m)\mathcal{U}_m \tag{7.12}$$

with

$$\hat{r}_\omega(m) := \int_I r_\omega(s)\mathcal{U}_m(s)w(s)\,ds$$

and

$$\sum_{m=n+1}^{\infty} |\hat{r}_\omega(m)|^2 = \|p_w - r_\omega\|^2_{L_2(I,w)} \le cn^{-2s}\|p_\omega\|^2_{W^s(L_2(I,w))}. \tag{7.13}$$

We define

$$R(\mathbf{x}) := \sum_{\omega\in\Omega_n} \lambda_\omega r_\omega(\mathbf{x}\cdot\omega)$$

which is an element of Y_N. Then we have, by (7.10) and (7.12),

$$R(\mathbf{x}) - P(\mathbf{x}) = \sum_{\omega\in\Omega_n}\sum_{m=n+1}^{\infty} \lambda_\omega\hat{r}_\omega(m)\mathcal{U}_m(\omega\cdot\mathbf{x}) = \sum_{m=n+1}^{\infty}\sum_{\omega\in\Omega_n} \lambda_\omega\hat{r}_\omega(m)\mathcal{U}_m(\omega\cdot\mathbf{x}).$$

We write

$$R_m(\mathbf{x}) := \sum_{\omega\in\Omega_n} \lambda_\omega\hat{r}_\omega(m)\mathcal{U}_m(\mathbf{x}\cdot\omega).$$

We have by Theorem 3.1 (see also (3.19)–(3.21))

$$R_m(\mathbf{x}) = \nu_m \int_{\mathbf{S}^{d-1}} A_m(R_m,\xi)\mathcal{U}_m(\xi\cdot\mathbf{x})\,d\xi,$$

where

$$\begin{aligned} A_m(R_m,\xi) &= \int_{\mathbf{B}^d} R_m(\mathbf{y})\mathcal{U}_m(\mathbf{y}\cdot\xi)\,d\mathbf{y} = \sum_{\omega\in\Omega_n}\lambda_\omega\hat{r}_\omega(m)\int_{\mathbf{B}^d}\mathcal{U}_m(\omega\cdot\mathbf{y})\mathcal{U}_m(\xi\cdot\mathbf{y})\,d\mathbf{y} \\ &= \sum_{\omega\in\Omega_n}\lambda_\omega\hat{r}_\omega(m)\frac{\mathcal{U}_m(\xi\cdot\omega)}{\mathcal{U}_m(1)}. \end{aligned}$$

We now use Theorem 3.1 and Lemma 4.4 to obtain

$$\begin{aligned} \|R_m\|^2_{L_2(\mathbf{B}^d)} &= \nu_m\|A_m(R_m,\omega)\|^2_{L_2(\mathbf{S}^{d-1})} = \nu_m^{-1}\left\|\sum_{\omega\in\Omega_n}\lambda_\omega\hat{r}_\omega(m)\frac{\nu_m}{\mathcal{U}_m(1)}\mathcal{U}_m(\xi\cdot\omega)\right\|^2_{L_2(\mathbf{S}^{d-1})} \\ &\le c\nu_m^{-1}(m/n)^{d-1}\sum_{\omega\in\Omega_n}\lambda_\omega|\hat{r}_\omega(m)|^2 \le cn^{-d+1}\sum_{\omega\in\Omega_n}\lambda_\omega|\hat{r}_\omega(m)|^2, \end{aligned}$$

where we used that $\nu_m \asymp m^{d-1}$ (see (3.8)). From this, (7.11), and (7.13), we find, using the Parseval identity (3.9),

$$\begin{aligned} \|R-P\|^2_{L_2(\mathbf{B}^d)} &= \sum_{m=n+1}^{\infty}\|R_m\|^2_{L_2(\mathbf{B}^d)} \le cn^{-d+1}\sum_{m=n+1}^{\infty}\sum_{\omega\in\Omega_n}\lambda_\omega|\hat{r}_\omega(m)|^2 \\ &= cn^{-d+1}\sum_{\omega\in\Omega_n}\lambda_\omega\sum_{m=n+1}^{\infty}|\hat{r}_\omega(m)|^2 \\ &\le cn^{-2s-d+1}\sum_{\omega\in\Omega_n}\|\lambda_\omega^{1/2}p_\omega\|^2_{W^s(L_2(I,w))} \\ &\le cn^{-2s-d+1}\|f\|^2_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))}. \end{aligned}$$

Thus there is a function $R \in Y_N$, $N = k_0 n$, such that (7.3) holds. Theorem 7.1 is now proved. □

REMARK 7.2. *As in* [DOP], *it is possible to prove Theorem* 7.1 *without using Lemma* 7.1. *In place of this lemma one uses a slightly stronger assumption than estimate* (7.1). *The corresponding proof would be more constructive than the present one. We do not provide the details of this approach but instead refer the reader to* [DOP].

8. Elimination of the weight w. The result of section 7 (Theorem 7.1) gives sufficient conditions on a sequence of univariate spaces X_n, $n = 1, 2, \dots$, in order that the spaces Y_n defined by (7.2) with Ω_n from (4.19) provide approximation rates for functions in Sobolev spaces $W^\alpha(L_2(\mathbf{B}^d))$ comparable with polynomials and splines. However, the assumption (7.1) imposed on X_n is inconvenient for direct application because of the appearance of the weight $w(t) := w_{d/2}(t) := (1-t^2)^{(d-1)/2}$. We shall show in this section how the weight factor w can be avoided so that the result of section 7 applies more directly. We shall consider approximation on the ball $\mathbf{B}^d_{1/2} := \{\mathbf{x} \in \mathbf{R}^d : |\mathbf{x}| \le 1/2\}$ rather than $\mathbf{B}^d$. Approximation on $\mathbf{B}^d$ or other balls follows by a change of variables.

We begin by assuming that we have in hand n-dimensional linear spaces Z_n of univariate functions defined on $J := [-1/2, 1/2]$ which satisfy a Jackson-type estimate similar to (7.1) but with weight $= 1$. Let $W^m(L_2(J))$, $m = 1, 2, \dots$, be the Sobolev space of functions $g \in L_2(J)$ such that $g^{(m)}$ is in $L_2(J)$. The seminorm and norm for $W^m(L_2(J))$ are defined by

$$|g|_{W^m(L_2(J))} := \|g^{(m)}\|_{L_2(J)} \quad ; \quad \|g\|_{W^m(L_2(J))} := \|g^{(m)}\|_{L_2(J)} + \|g\|_{L_2(J)}.$$

For $0 < s < m$ not an integer, we define $W^s(L_2(J))$ by interpolation:

$$W^s(L_2(J)) := (L_2(J), W^m(L_2(J)))_{\theta,2}, \quad \theta := s/m, \tag{8.1}$$

with the norm as the interpolation space norm. For a given value of s, different values of $m > s$ give equivalent norms (see [DL]).

Our assumption on Z_n is that for a certain fixed value of s, we have that for each $g \in W^s(L_2(J))$, there is a function $\zeta_n \in Z_n$ such that

$$\|g - \zeta_n\|_{L_2(J)} \le c(s) n^{-s} \|g\|_{W^s(L_2(J))} \tag{8.2}$$

with the constant $c(s)$ depending only on s.

Let X_n be the space of univariate functions r such that for some $p \in \mathcal{P}_n(I)$ and some $\zeta \in Z_n$,

$$r(t) = \begin{cases} p(t), & t \in I \setminus J, \\ \zeta(t), & t \in J. \end{cases} \tag{8.3}$$

We shall show that under the assumption (8.2) on the Z_n, the spaces X_n, $n = 1, 2, \dots$, satisfy the assumption (7.1). To prove this, we recall the definition (6.4) of the spaces $W^\alpha(L_2(I, w))$ and the operator Λ of (2.12):

$$\Lambda g := \left(\frac{d}{dt}\right)^{d-1} [wg]\,. \tag{8.4}$$

According to (2.15), we have $\Lambda^2 \mathcal{U}_n = (-1)^{d-1} \mu_n^2 \mathcal{U}_n$. Since $\mu_n \asymp n^{d-1}$ (see (2.16)), it follows that for each $g \in W^{m\lambda}(L_2(I, w))$, $\lambda = 2(d-1)$, $m = 1, 2, \dots$, we have

$$\|g\|_{W^{m\lambda}(L_2(I,w))} \asymp \|\Lambda^{2m} g\|_{L_2(I,w)} \tag{8.5}$$

with the constants of equivalency depending only on d.

LEMMA 8.1. *For each $m = k\lambda$, with $\lambda := 2(d-1)$ and k a nonnegative integer, we have*

$$\|g\|_{W^m(L_2(J))} \le c(d,m)\|g\|_{W^m(L_2(I,w))} \quad \text{for } g \in W^m(L_2(I,w)) \tag{8.6}$$

with the constant $c(d,m)$ depending only on d and m.

Proof. We first observe that the weight w is strictly positive on J and, therefore, w^{-1} is infinitely times differentiable on J. Then the following identity holds:

$$g^{(\ell(d-1))} = \sum_{j=0}^{\ell(d-1)-1} u_j g^{(j)} + u_{\ell(d-1)}\Lambda^\ell g, \quad \ell = 1,2,\dots, \tag{8.7}$$

where u_j are obtained from w^{-1} and its derivatives. Indeed, (8.7) can be proved by induction on ℓ. For $\ell = 1$, (8.7) follows from Leibniz's formula for differentiating the product $g = w^{-1}(wg)$. Suppose that (8.7) holds for some $\ell \ge 1$. Then one writes $\Lambda^\ell g$ as $w^{-1}\left(w\Lambda^\ell g\right)$ and differentiates both sides of (8.7) $d-1$ times to prove it for $\ell+1$.

It follows from (8.7), with $\ell = 2k$ and $m = k\lambda$, that

$$\|g^{(m)}\|_{L_2(J)} \le c\sum_{j=0}^{m-1} \|g^{(j)}\|_{L_2(J)} + c\|\Lambda^k g\|_{L_2(J)}. \tag{8.8}$$

We shall use next the following well-known inequality (see, e.g., [BS])

$$\|g^{(j)}\|_{L_2(J)} \le c\left(\delta^{-j}\|g\|_{L_2(J)} + \delta^{m-j}\|g^{(m)}\|_{L_2(J)}\right), \quad j = 1,2,\dots,m, \tag{8.9}$$

where $\delta > 0$ is arbitrary and c depends only on m. Combining (8.8) with (8.9) we get, for $0 < \delta < 1$,

$$\|g^{(m)}\|_{L_2(J)} \le c^*\delta^{-m+1}\|g\|_{L_2(J)} + c^*\delta\|g^{(m)}\|_{L_2(J)} + c^*\|\Lambda^k g\|_{L_2(J)}, \tag{8.10}$$

where $c^* > 1$ is independent of δ. We now select δ such that $c^*\delta = 1/2$ and bring the second term on the right in (8.10) to the left-hand side. We obtain

$$\begin{aligned}\|g^{(m)}\|_{L_2(J)} &\le c(\|g\|_{L_2(J)} + \|\Lambda^k g\|_{L_2(J)})\\ &\le c(\|wg\|_{L_2(I)} + \|w\Lambda^k g\|_{L_2(I)})\\ &\le c\|g\|_{W^m(L_2(I,w))}. \qquad \square\end{aligned}$$

THEOREM 8.1. *If the sequence of spaces Z_n, $n = 1,2,\dots$, satisfies* (8.2)*, then the spaces X_n, $n = 1,2,\dots$, defined by* (8.3) *satisfy the Jackson estimates* (7.1)*; i.e., for each univariate function $g \in W^s(L_2(I,w))$, there is a function $r \in X_n$ which provides the Jackson estimate*

$$\|g - r\|_{L_2(I,w)} \le cn^{-s}\|g\|_{W^s(L_2(I,w))}, \tag{8.11}$$

with c a constant independent of g and n.

Proof. Consider the linear operator T that associates with every function $g \in L_2(I,w)$ the restriction of g on J. Since w is strictly positive on J, T is a bounded operator from $L_2(I,w)$ into $L_2(J)$. By Lemma 8.1, T is bounded from $W^m(L_2(I,w))$ into $W^m(L_2(J))$ for each $m = 2k(d-1)$, $k = 1,2,\dots$. This implies that, for each

$0 < s \le 2k(d-1)$ and $\theta := s/(2k(d-1))$, we have by interpolation (see (6.8) and (8.1)) that for each $g \in W^s(L_2(I,w))$,

$$\begin{aligned}\|g\|_{W^s(L_2(J))} &\asymp \|g\|_{\left(L_2(J),W^{2k(d-1)}(L_2(J))\right)_{\theta,2}} \\ &\le c\|g\|_{\left(L_2(I,w),W^{2k(d-1)}(L_2(I,w))\right)_{\theta,2}} \asymp \|g\|_{W^s(L_2(I,w))}.\end{aligned}$$

Now, given $g \in W^s(L_2(I,w))$, we let $\zeta \in Z_n$ satisfy (8.2). Then, from (8.6),

$$\|g-\zeta\|_{L_2(J)} \le cn^{-s}\|g\|_{W^s(L_2(J))} \le cn^{-s}\|g\|_{W^s(L_2(I,w))}.$$

Similarly, let p be the best approximation in $L_2(I,w)$ to g from $\mathcal{P}_n(I)$. Then, from (6.5),

$$\|g-p\|_{L_2(I,w)} \le n^{-s}\|g\|_{W^s(L_2(I,w))}.$$

It now follows that the function $r \in X_n$ defined by (8.3) for these ζ and p satisfies (8.11). □

THEOREM 8.2. *If the sequence of spaces $Z_n, n = 1,2,\dots,$ satisfy* (8.2)*, then for any function $f \in W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))$, there are functions $r_\omega \in Z_n$ such that*

$$R(\mathbf{x}) = \sum_{\omega\in\Omega_n} r_\omega(\omega\cdot\mathbf{x}) \tag{8.12}$$

satisfies

$$\|f-R\|_{L_2(\mathbf{B}^d_{1/2})} \le cn^{-s-(d-1)/2}\|f\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))} \tag{8.13}$$

with c independent of f and n.

Proof. We first recall (see, e.g., [A, Chapter IV]) that f can be extended to a function f_0 defined on all of $\mathbf{R}^d$ such that f_0 vanishes outside of $\mathbf{B}^d_{3/4}$ and

$$\|f_0\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d))} \le c\|f\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))}$$

with a constant c depending only on s and d.

We define X_n as in (8.3). From Theorem 8.1, we obtain that condition (7.1) is satisfied. Therefore, from Theorem 7.1 there are functions $r_\omega \in X_n$, $\omega \in \Omega_n$, such that the function

$$R(\mathbf{x}) = \sum_{\omega\in\Omega_n} r_\omega(\omega\cdot\mathbf{x})$$

satisfies

$$\begin{aligned}\|f_0-R\|_{L_2(\mathbf{B}^d)} &\le cn^{-r-(d-1)/2}\|f_0\|_{W^{r+(d-1)/2}(L_2(\mathbf{B}^d))} \\ &\le cn^{-r-(d-1)/2}\|f\|_{W^{r+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))}.\end{aligned} \tag{8.14}$$

On the ball $\mathbf{B}^d_{1/2}$, $f_0 = f$ and r_ω is in Z_n for each $\omega \in \Omega_n$. Therefore, (8.13) follows from (8.14). □

9. Examples and further remarks. In this section, we shall give some applications of the results of section 8. Theorem 8.2 implies that for any sequence of spaces Z_n, $n = 1, 2, \ldots$, contained in $L_2(J)$, $J = [-1/2, 1/2]$, that satisfy (8.2) we have the estimate (8.13) for $f \in W^{s+(d-1)/2}(L_2(J))$. The condition (8.2) is satisfied by all the standard spaces of approximation such as algebraic polynomials and spline functions (discussed in more detail later in this section). We wish to single out, for further elaboration, one particular example which appears frequently in wavelet theory, as well as in computer aided design.

Let ϕ be a univariate function with compact support on $\mathbf{R}$. Let ℓ be the smallest integer such that ϕ or one of its shifts $\phi(x-k)$, $k \in \mathbf{Z}$, is supported on $[0, \ell]$. If necessary, we can redefine ϕ to be one of its integer shifts and thereby require that ϕ is supported on $[0, \ell]$. We denote by $\mathcal{S} := \mathcal{S}(\phi)$ the shift-invariant space which is the $L_2(\mathbf{R})$-closure of finite linear combinations of the shifts $\phi(\cdot - j)$, $j \in \mathbf{Z}$, of ϕ. By dilation, we obtain the univariate spaces

$$\mathcal{S}^k := \{S(2^k \cdot) : S \in \mathcal{S}\}, \quad k \in \mathbf{Z}.$$

The approximation properties of the family of spaces $\mathcal{S}^k$ is well understood. In [BDR], there is a complete characterization (in terms of the Fourier transform of ϕ) of when the spaces $\mathcal{S}^k$ provide the Jackson estimates

$$\operatorname{dist}(g, \mathcal{S}^k)_{L_2(\mathbf{R})} \le C 2^{-ks} \|g\|_{W^s(L_2(\mathbf{R}))}. \tag{9.1}$$

For an integer s, we say that ϕ satisfies the Strang–Fix conditions of order s if

$$\hat{\phi}(0) \neq 0, \text{ and } D^j \hat{\phi}(2k\pi) = 0, \ k \in \mathbf{Z}, \ k \neq 0, \ j = 0, 1, \ldots, s-1. \tag{9.2}$$

If ϕ satisfies (9.2) and ϕ is piecewise continuous and of bounded variation, then $\mathcal{S}^k$ provides the approximation estimate (9.1) (see, e.g., [DL, Chapter 13]).

We denote by $\mathcal{S}^k(J)$, $k \ge 1$, the restrictions of the spaces $\mathcal{S}^k$ to the interval $J := [-1/2, 1/2]$. The functions $\phi(2^k t - j)$, $j = -\ell + 1 - 2^{k-1}, \ldots, 2^{k-1} - 1$, span $\mathcal{S}^k(J)$. Each function g in $W^s(L_2(J))$ can be extended to $\mathbf{R}$ with

$$\|g\|_{W^s(L_2(\mathbf{R}))} \le c \|g\|_{W^s(L_2(J))}.$$

It follows therefore that the spaces $\mathcal{S}^k(J)$ provide the approximation property (8.2) and hence Theorem 8.2 applies with $n = 2^k$. The functions R appearing in Theorem 8.2 are of the form

$$R(\mathbf{x}) = \sum_{j=-\ell+1-2^{k-1}}^{2^k-1} \sum_{\omega \in \Omega_{2^k}} c(j, \omega) \phi(2^k \mathbf{x} \cdot \omega - j).$$

There is another representation of the functions in $\mathcal{S}^k(J)$ related to sigmoidal functions. Let

$$\sigma(t) := \sum_{j=0}^{\infty} \phi(t - j). \tag{9.3}$$

Then the functions $\sigma(2^k t - j)$, $j = -\ell + 1 - 2^{k-1}, \ldots, 2^{k-1} - 1$, also span $\mathcal{S}^k(J)$. The function σ is 0 for t sufficiently large negative and 1 for t sufficiently large positive. However, it is not necessarily monotone (without additional assumptions on ϕ).

COROLLARY 9.1. *Let ϕ satisfy the Strang–Fix conditions* (9.2) *of order s. Then for each function $f \in W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))$, there is a function*

$$R(\mathbf{x}) = \sum_{j=-\ell+1+2^{k-1}}^{2^{k-1}-1} \sum_{\omega \in \Omega_{2^k}} c(j,\omega)\sigma(2^k \mathbf{x}\cdot\omega - j)$$

such that

$$\|f - R\|_{L_2(\mathbf{B}^d_{1/2})} \le c\, 2^{-(s+(d-1)/2)k} \|f\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))}, \quad k = 1, 2, \dots,$$

with c independent of f and k.

For certain choices of ϕ above, we obtain that σ of (9.3) is a sigmoidal function in the terminology of neural networks. We recall that a sigmoidal function is a nonnegative, monotone, univariate function which has limits $= 0$ as $t \to -\infty$ and $= 1$ as $t \to \infty$. To obtain examples of such sigmoidal functions, we can take ϕ to be a B-spline. Let $\phi := N_{0,s}$, where for each $j \in \mathbf{Z}$ and $s = 1, 2, \dots$, $N_{j,s} := s^{-1} M_{j,s}$ is the B-spline of order s (see [DL, Chapter 5]) with breakpoints $\frac{j}{2n}, \dots, \frac{j+s}{2n}$. The function

$$\sigma_s(t) := \sum_{j=-s+1}^{\infty} N_{j,s}(t), \quad t \in \mathbf{R}$$

of (9.3) is a sigmoidal function, and, in the case $s = 1$, it is the unit impulse function $\chi_{[0,\infty)}$. The functions $\sigma_s(t - \frac{j}{2n})$, $j = -n, \dots, n+s-1$, form a basis for $\mathcal{S}_{n,s}$ the space of all splines of degree $s-1$ defined on J with breakpoints belonging to the set $\{\frac{-n+1}{2n}, \frac{-n+2}{2n}, \dots, \frac{n-1}{2n}\}$. From Theorem 8.2, we obtain the following.

COROLLARY 9.2. *For any $f \in W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))$, there are constants $c(k,\omega)$, $\omega \in \Omega_n$, $k = -n, \dots, n+s-1$, such that*

$$R(\mathbf{x}) = \sum_{\omega \in \Omega_n} \sum_{k=-n}^{n+s-1} c(k,\omega)\sigma_s\left(\mathbf{x}\cdot\omega - \frac{k}{2n}\right) \tag{9.4}$$

satisfies

$$\|f - R\|_{L_2(\mathbf{B}^d_{1/2})} \le c n^{-s-(d-1)/2} \|f\|_{W^{s+(d-1)/2}(L_2(\mathbf{B}^d_{1/2}))}$$

with c independent of f and n.

The functions R in (9.4) correspond to the outputs of a feed-forward neural network with $O(n^{d-1})$ nodes of computation. Thus, the corollary shows that such neural networks have computational efficiency comparable with standard methods of approximation like splines and wavelets.

The special case $s = 1$ in Corollary 9.2 is also noteworthy. In this case the function σ is the unit-impulse function and the functions R are piecewise constant. The order of approximation provided by Corollary 9.2 is somewhat surprising. One might expect that such piecewise constants could only provide approximation order 1 while the corollary gives approximation order $(d+1)/2$.

10. Appendix.

A1. Proof of (3.4). Since $\mathcal{P}_n$ is invariant under rotations, it is sufficient to prove that $\langle P(\mathbf{x}),\mathcal{U}_n(x_1)\rangle = 0$ for each $P \in \mathcal{P}_{n-1}$ or that

$$\langle \mathbf{x}^{\mathbf{m}},\mathcal{U}_n(x_1)\rangle := \int_{\mathbf{B}^d} \mathbf{x}^{\mathbf{m}}\mathcal{U}_n(x_1)\,d\mathbf{x} = 0 \quad \text{when} \quad |\mathbf{m}| \le n-1.$$

Write

$$\mathbf{B}_{x_1} := \{\mathbf{x}' = (x_2,\dots,x_d) : x_2^2 + \cdots + x_d^2 \le 1 - x_1^2\}.$$

We have

$$\langle \mathbf{x}^{\mathbf{m}},\mathcal{U}_n(x_1)\rangle = \int_I x_1^{m_1}\left(\int_{\mathbf{B}_{x_1}} x_2^{m_2}\dots x_d^{m_d}\,d\mathbf{x}'\right)\mathcal{U}_n(x_1)\,dx_1.$$

Because of the symmetry it is obvious that the inner integral above is equal to zero if at least one of $m_2,\dots,m_d$ is odd. Consider the case when all $m_2,\dots,m_d$ are even. We now change the rectangular coordinates in the inner integral to spherical and find

$$\int_{\mathbf{B}_{x_1}} x_2^{m_2}\dots x_d^{m_d}\,d\mathbf{x}' = c\int_0^{(1-x_1^2)^{1/2}} r^{m_2+\cdots+m_d+d-2}\,dr = c(1-x_1^2)^{\frac{1}{2}(m_2+\cdots+m_d+d-1)},$$

where c depends on $d, m_2,\dots,m_d$. Therefore,

$$\langle \mathbf{x}^m, C_n(x_1)\rangle = c\int_I x_1^{m_1}(1-x_1^2)^{\frac{1}{2}(m_2+\cdots+m_d)}\mathcal{U}_n(x_1)(1-x_1^2)^{\frac{d-1}{2}}\,dx_1 = 0$$

since the univariate polynomial $\mathcal{U}_n$ is orthogonal to $\mathcal{P}_{n-1}(I)$ in $L_2(I,w)$ (see (3.3) and (2.1)). □

A2. Proof of (3.10). We first show that for each $g \in L_1(I, w_{(d-1)/2})$

$$\mathcal{R}(g(\eta\cdot\mathbf{x});\xi,t) = |B^{d-2}|(1-t^2)^{\frac{d-1}{2}}\int_I g(\cos\theta\cos\psi + u\sin\theta\sin\psi)(1-u^2)^{\frac{d-2}{2}}\,du, \tag{10.1}$$

where $t =: \cos\theta$, $t \in I$, $\psi \in [0,\pi]$ is the angle between ξ and η ($\cos\psi = \xi\cdot\eta$), $|B^{d-2}|$ is the volume of the unit ball B^{d-2} in $\mathbf{R}^{d-2}$, $|B^{d-2}| = \frac{2\pi^{d/2-1}}{(d-2)\Gamma(d/2-1)}$, and $\mathcal{R}$ is the Radon transform defined in (3.23). Indeed, it is easily seen that

$$\mathcal{R}(g(\eta\cdot\mathbf{x});\xi,t) = |B^{d-2}|\int_{-\sqrt{1-t^2}}^{\sqrt{1-t^2}} g(t\cos\psi + v\sin\psi)(1-t^2-v^2)^{\frac{d-2}{2}}\,dv.$$

Substituting $v = (1-t^2)^{1/2}u$ in the above integral we get (10.1).

Our second step is to prove that

$$\mathcal{R}(C_n^{d/2}(\eta\cdot\mathbf{x});\xi,t) = |B^{d-2}|\frac{2^{d-1}\Gamma^2(d/2)n!}{\Gamma(n+d)}(1-t^2)^{\frac{d-1}{2}}C_n^{d/2}(t)C_n^{d/2}(\xi\cdot\eta). \tag{10.2}$$

Indeed, the classical addition theorem for Legendre (Gegenbauer) polynomials can be written as follows (see [E, p. 178]):

$$\begin{aligned} &C_n^\lambda(\cos\theta\cos\psi + \sin\theta\sin\psi\cos\varphi)\\ &= \sum_{m=0}^n 2^m(2\lambda+2m-1)(n-m)!\frac{[(\lambda)_m]^2}{(2\lambda-1)_{n+m+1}}\\ &\times (\sin\theta)^m C_{n-m}^{\lambda+m}(\cos\theta)(\sin\psi)^m C_{n-m}^{\lambda+m}(\cos\psi)C_m^{\lambda-1/2}(\cos\varphi)\end{aligned}$$

and hence, for $\lambda = d/2$, and $u := \cos\varphi$, $u \in I$, we have

$$\begin{aligned} &C_n^{d/2}(\cos\theta\cos\psi + u\sin\theta\sin\psi) \\ &= \sum_{m=0}^{n} 2^m(d+2m-1)(n-m)!\frac{[(d/2)_m]^2}{(d-1)_{n+m+1}} \\ &\times (\sin\theta)^m C_{n-m}^{d/2+m}(\cos\theta)(\sin\psi)^m C_{n-m}^{d/2+m}(\cos\psi)C_m^{(d-1)/2}(u). \end{aligned}$$

We now insert this into (10.1) and use the fact that $C_m^{(d-1)/2}(u)$, $m = 1, 2, \ldots$, are orthogonal to the constants in $L_2(I, w_{(d-1)/2})$ to obtain

$$\begin{aligned} &\mathcal{R}(C_n^{d/2}(\eta\cdot\mathbf{x});\xi,t) \\ &= |B^{d-2}|(1-t^2)^{(d-1)/2}\frac{(d-1)n!}{(d-1)_{n+1}}C_n^{d/2}(\cos\theta)C_n^{d/2}(\cos\psi)\int_I (1-u^2)^{(d-2)/2}\,du. \end{aligned}$$

This implies (10.2).

We finally use (10.2) to obtain

$$\begin{aligned} &\int_{\mathbf{B}^d} C_n^{d/2}(\eta\cdot\mathbf{x})C_n^{d/2}(\xi\cdot\mathbf{x})\,d\mathbf{x} = \int_I \mathcal{R}(C_n^{d/2}(\eta\cdot\mathbf{x});\xi,t)C_n^{d/2}(t)\,dt \\ &= |B^{d-2}|\frac{2^{d-1}\Gamma^2(d/2)n!}{\Gamma(n+d)}C_n^{d/2}(\eta\cdot\xi)\int_I [C_n^{d/2}(t)]^2(1-t^2)^{(d-1)/2}\,dt \\ &= \gamma_{n,d}C_n^{d/2}(\eta\cdot\xi), \end{aligned}$$

where

$$\gamma_{n,d} := |B^{d-2}|\frac{2^{d-1}\Gamma^2(\frac{d}{2})n!}{\Gamma(n+d)}h_{n,d/2}.$$

Simple calculations show that this is (3.10). See [RK]. □

A3. Proof of (3.16). The following relation between contiguous Gegenbauer polynomials holds (see [E, p. 178], (36)):

$$(n+\lambda)C_{n+1}^{\lambda-1} = (\lambda-1)\left[C_{n+1}^{\lambda} - C_{n-1}^{\lambda}\right], \quad \lambda > 1.$$

Also, $C_0^\lambda(t) = 1$ and $C_1^\lambda(t) = 2\lambda t$. These identities readily imply

$$C_n^\lambda = \sum_{j=0}^{[n/2]} \frac{n-2j+\lambda-1}{\lambda-1}C_{n-2j}^{\lambda-1}. \tag{10.3}$$

Simple calculations show that (10.3) with $\lambda = d/2$ ($d > 2$) is (3.16). □

A4. Proof of (3.11). Identity (3.11) follows from the fact that $\mathcal{U}_n(\xi\cdot\mathbf{x})$ (as a function of ξ) is a spherical polynomial in $\mathcal{H}_n \oplus \mathcal{H}_{n-2} \oplus \cdots \oplus \mathcal{H}_\epsilon$ and $\nu_n\mathcal{U}_n(\xi\cdot\eta)/\mathcal{U}_n(1)$ is the reproducing kernel for this space (see (3.17)). □

Acknowledgments. The author is enormously grateful to Ronald DeVore, who read an early version of this paper and made many suggestions for improvements. The author is also deeply grateful to Konstantin Oskolkov for various helpful discussions.

REFERENCES

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

[BDR] C. de Boor, R. DeVore, and A. Ron, *Approximation from shift invariant spaces*, Trans. Amer. Math. Soc., 341 (1994), pp. 787–806.

[BS] C. Bennett and R. Sharpley, *Interpolation of Operators*, Academic Press, New York, 1988.

[DL] R. DeVore and G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer, Heidelberg, 1993.

[DOP] R. DeVore, K. Oskolkov, and P. Petrushev, *Approximation by feed-forward neural networks*, Ann. Numer. Math., 4 (1997), pp. 261–287.

[E] A. Erdelyi, ed., *Higher Transcendental Functions*, Vol. 2, McGraw–Hill, New York, 1953.

[K] E. Kogbetliantz, *Recherches sur la sommabilité des séries ultraspheriques par la méthode de moyennes arithmetiques*, J. Math. Pures Appl., 9 (1924), pp. 107–187.

[L] D. Ludwig, *The Radon transform on Euclidean spaces*, Comm. Pure Appl. Math., 19 (1966), pp. 49–81.

[LS] B. Logan and L. Schepp, *Optimal reconstruction of a function from its projections*, Duke Math. J., 42 (1975), pp. 645–659.

[M] H. Mhaskar, *Neural networks for optimal approximation of smooth and analytic functions*, Neural Comput., 8 (1996), pp. 164–177.

[MM] H. Mhaskar and C. Micchelli, *Approximation by superposition of sigmoidal and radial basis functions*, Adv. Appl. Math., 13 (1992), pp. 350–373.

[MM1] H. Mhaskar and C. Micchelli, *Degree of approximation by neural and translation networks with a single hidden layer*, Adv. Appl. Math., 16 (1995), pp. 151–183.

[O] K. Oskolkov, *Inequalities of the "large sieve" type and applications to problems of trigonometric approximation*, Anal. Math., 12 (1986), pp. 143–166.

[P] S. Pawelke, *Über die Approximationsordnung bei Kugelfunktionen und algebraischen Polynomen*, Tôhoku Math. J., 24 (1972), pp. 473–486.

[RK] A. Ramm and A. Katsevich, *The Radon Transform and Local Tomography*, CRC Press, Boca Raton, FL, 1996.

[Se] R. Seeley, *Spherical harmonics*, Amer. Math. Monthly, 73 (1966), pp. 115–121.

[St] E. Stein, *Interpolation in polynomial classes and Markoff's inequality*, Duke Math. J., 24 (1957), pp. 467–476.

[SW] E. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.

[Sz] G. Szegö, *Orthogonal Polynomials*, American Mathematical Society, Providence, RI, 1975.

[Z] A. Zygmund, *Trigonometric Series*, Vols. I, II, Cambridge University Press, Cambridge, 1977.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 190–219

BULK AND CONTACT ENERGIES: NUCLEATION AND RELAXATION*

IRENE FONSECA† AND GIOVANNI LEONI‡

Abstract. An integral representation formula in $BV(\Omega;\mathbb{R}^p)$ for the relaxation $\mathcal{H}(u,\Omega)$ with respect to the L^1 topology of functionals of the general form

$$H(u,\Omega) := \int_\Omega h(x,u(x),\nabla u(x))\,dx + \int_{\partial\Omega} \theta(x,T\,u(x))\,dH_{N-1}(x), \quad u \in W^{1,1}(\Omega;\mathbb{R}^p),$$

is obtained. Here $\Omega \subset \mathbb{R}^N$ is an open, bounded set of class C^2, T is the trace operator on $\partial\Omega$, and H_{N-1} is the $N-1$-dimensional Hausdorff measure. The main hypotheses on the functions h and θ are that $h(x,u,\cdot)$ is quasiconvex and has linear growth, and that $\theta(x,\cdot)$ is Lipschitz. The understanding of nucleation phenomena for materials undergoing phase transitions leads to the study of constrained minimization problems of the type

$$\inf\left\{\mathcal{H}(u,\Omega) + \int_\Omega \tau(x,u(x))\,dx :\ u \in BV(\Omega;K)\right\},$$

where K is a nonempty compact subset of $\mathbb{R}^p$, and $\tau : \Omega \times K \to \mathbb{R}$ is a continuous function. It is shown that if $\tau(x,\cdot)$ is a double well potential vanishing only at α and β, then minimizers u of the total energy are given by pure phases; that is, there exists $\Omega_u \subset \Omega$ such that $u(x) = \alpha$ for $\mathcal{L}^N$ a.e. $x \in \Omega_u$ (liquid) and $u(x) = \beta$ for $\mathcal{L}^N$ a.e. $x \in \Omega\backslash\Omega_u$ (solid). This conclusion is closely related to results previously obtained by Visintin, and where the interfacial energy is assumed to satisfy a *generalized co-area formula*. Here this condition is replaced by a property which may be verified by energies for which the co-area formula might not hold.

Key words. functions of bounded variation, nucleation, relaxation, bulk and contact energies, generalized co-area formula

AMS subject classifications. 49J45, 49Q20, 49N60, 73T05, 73V30

PII. S0036141097325563

1. Introduction.

This paper is divided into two parts. In the first part we obtain an integral representation formula in $BV(\Omega;\mathbb{R}^p)$ for the relaxation $\mathcal{H}(u,\Omega)$ with respect to the L^1 topology of functionals of the general form

(1.1)
$$H(u,\Omega) := \int_\Omega h(x,u(x),\nabla u(x))\,dx + \int_{\partial\Omega} \theta(x,T\,u(x))\,dH_{N-1}(x), \quad u \in W^{1,1}(\Omega;\mathbb{R}^p),$$

where $\Omega \subset \mathbb{R}^N$ is an open, bounded set of class C^2, T is the trace operator on $\partial\Omega$, and H_{N-1} is the $N-1$-dimensional Hausdorff measure. The main hypotheses on the functions h and θ are that $h(x,u,\cdot)$ is quasiconvex and has linear growth, and that $\theta(x,\cdot)$ is Lipschitz.

*Received by the editors August 6, 1997; accepted for publication (in revised form) March 19, 1998; published electronically October 26, 1998.
http://www.siam.org/journals/sima/30-1/32556.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (fonseca@andrew.cmu.edu). The research of this author was partially supported by the National Science Foundation through the Center for Nonlinear Analysis, and by the National Science Foundation under grant DMS–9500531.

‡Center for Nonlinear Analysis, Carnegie Mellon University, Pittsburgh, PA 15213, and Department of Mathematics, University of Perugia, Perugia, Italy 06123 (leoni@dipmat.unipg.it). Part of this research was undertaken during the author's visit to the Center for Nonlinear Analysis under the CNA/CMU-CNR project.

Under a *degenerate* coercivity assumption on $h(x,u,\cdot)$ we obtain the following integral representation for $u \in BV(\Omega;\mathbb{R}^p)$:

$$\begin{aligned}\mathcal{H}(u,\Omega) = &\int_\Omega h(x,u(x),\nabla u(x))\,dx + \int_\Omega h^\infty(x,u(x),dC(u))\\ &+\int_{S(u)\cap\Omega} K_h(x,u^-(x),u^+(x),\nu_u(x))\,dH_{N-1}(x)\\ &+\int_{\partial\Omega}\theta(x,T\,u(x))\,dH_{N-1}(x),\end{aligned} \tag{1.2}$$

where ∇u is the density of the absolutely continuous part of the distributional derivative Du with respect to the N-dimensional Lebesgue measure $\mathcal{L}^N$, $(u^+ - u^-)$ is the jump across the interface $S(u)$, and $C(u)$ is the Cantor part of Du. For the canonical model where $h(x,u,\nabla u) := \sigma|\nabla u|$, $\sigma > 0$, the relaxed energy $\mathcal{H}(u,\Omega)$ reduces to

$$\mathcal{H}(u,\Omega) = \sigma\int_\Omega |Du| + \int_{\partial\Omega}\theta(x,T\,u)\,dH_{N-1}, \qquad u \in BV(\Omega;\mathbb{R}^p). \tag{1.3}$$

In the scalar case where $p = 1$, the lower semicontinuity of the functional (1.3) was proved by Massari and Pepe [MP] when $\theta(x,u) := \hat\sigma|u|$, with $|\hat\sigma| \le \sigma$, and by Modica [Mo2] under the assumption that

$$|\theta(x,u) - \theta(x,u_1)| \le \sigma|u - u_1| \tag{1.4}$$

for all $x \in \partial\Omega$ and all u, $u_1 \in \mathbb{R}$.

One of the motivations for the introduction of a relaxed energy is that nonconvex variational problems may not have a minimizer in the space of smooth functions; therefore, in order to apply the direct method of calculus of variations one has to extend the original functional. Although Sobolev spaces are considered to be the natural extension to the space of smooth functions, in recent years the theory of phase transitions, and the need to determine effective energies for materials exhibiting instabilities such as fractures and defects, have led us to further extend the domain of functionals of the form (1.1) in order to include functions u which present discontinuities along surfaces. Motivated somewhat by Lebesgue's definition of surface area, Serrin in [Se1, Se2] proposed the following notion for the relaxed energy of $H(u,\Omega)$ (in the case where $\theta \equiv 0$):

$$\mathcal{H}(u,\Omega) := \inf_{\{u_n\}}\left\{\liminf_{n\to\infty} H(u_n,\Omega) :\ u_n \in W^{1,1}(\Omega;\mathbb{R}^p),\ u_n \to u \quad \text{in} \quad L^1(\Omega;\mathbb{R}^p)\right\}.$$

One of the main issues in the calculus of variations concerns the search and characterization of an integral representation for $\mathcal{H}(u,\Omega)$ in the space $BV(\Omega;\mathbb{R}^p)$.

In the scalar case where $p = 1$ and $h(x,u,\cdot)$ is convex, the integral representation (1.2) was first obtained by Goffman and Serrin [GSe] when $h = h(\nabla u)$ (see also [Re]), and by Giaquinta, Modica, and Souček [GMS] for $h = h(x,\nabla u)$. These results were then extended by Dal Maso [DM] who considered the general case where $h = h(x,u,\nabla u)$ and emphasized the important role of the coercivity condition in establishing (1.2). Indeed, Dal Maso showed that, while (1.2) holds for nonnegative functions $h = h(u,\nabla u)$ without any lower bound on h, when $h = h(x,\nabla u)$, or, more generally, when $h = h(x,u,\nabla u)$, the representation (1.2) may fail unless one requires a weak coercivity assumption of the form

$$h(x,u,\nabla u) \ge g(x,u)|\nabla u|. \tag{1.5}$$

In the vectorial case where $p > 1$ and $h(x, u, \cdot)$ is quasiconvex, Ambrosio and Dal Maso [ADM2] proved (1.2) when $h = h(\nabla u)$ and without (1.5). Independently, Fonseca and Müller [FM2] obtained this result for general functions $h(x, u, \nabla u)$ which verify (1.5).

In all the works mentioned above $\theta \equiv 0$, and one of the purposes of this paper is to extend these results to the new case where possibly $\theta \not\equiv 0$. The relaxation of functionals of the type (1.1) arises in the van der Waals–Cahn–Hilliard theory of phase transitions for fluids (cf. [vdW, CH1, CH 2]). In this context the boundary term $\int_{\partial\Omega} \theta(x, T\,u)\, dH_{N-1}$ represents the contact energy between the fluid and the container walls, where $\theta(x, u)$ is the contact energy per unit area when the density is u (see [C, G, Mo1]).

We present here two relaxation results. In Theorem 3.5 we show that, without any a priori coercivity on the function h, the functional on the right-hand side of (1.2) actually gives the integral representation for the following relaxed energy:

$$\mathcal{H}_b(u,\Omega) = \inf_{\{u_n\}} \Big\{ \liminf_{n\to\infty} H(u_n,\Omega) :\ u_n \in W^{1,1}(\Omega;\mathbb{R}^p),\ u_n \to u \text{ in } L^1(\Omega;\mathbb{R}^p),\ \sup_n ||u_n||_{W^{1,1}} < \infty \Big\},$$

while in Theorem 3.2 we prove that $\mathcal{H}_b(u, \Omega) = \mathcal{H}(u, \Omega)$ if h satisfies a condition of the type (1.5). Therefore we may conclude that the right-hand side of (1.2) always coincides with $\mathcal{H}_b(u, \Omega)$, and we restate all the results mentioned above by saying that $\mathcal{H}_b(u, \Omega) = \mathcal{H}(u, \Omega)$ in the scalar case if either $h = h(u, \nabla u)$ or if $h = h(x, u, \nabla u)$ satisfies (1.5), and in the vectorial case if either $h = h(\nabla u)$ or if $h = h(x, u, \nabla u)$ satisfies (1.5). In the remaining cases it may happen that $\mathcal{H}(u, \Omega) < \mathcal{H}_b(u, \Omega)$.

It is worth mentioning that the fact that the relaxation $\mathcal{H}(u, \Omega)$ is simply given by the decoupled sum of the relaxation of the functional $\int_\Omega h(x, u, \nabla u)\, dx$ and the contact energy may be somewhat deceiving, since it hides the competition between the bulk energy and the contact energy. A more insightful way to look at (1.1), and consequently, at (1.2), is perhaps to consider the equivalent form

$$\begin{aligned} H(u,\Omega) = &\int_\Omega \big\{ h(x,u(x),\nabla u(x)) + \varphi(x)\cdot \nabla u^T(x) \nabla_u \theta(x,u(x)) \big\}\, dx \\ &+ \int_\Omega \theta(x,u(x)) \operatorname{div}\varphi(x)\, dx + \int_\Omega \varphi(x)\cdot \nabla_x \theta(x,u(x))\, dx, \end{aligned} \tag{1.6}$$

where $\varphi \in C^1(\mathbb{R}^N;\mathbb{R}^N)$ depends only on Ω and $|\varphi(x)| < 1$ in Ω (see Lemma 2.1). In particular, in the isotropic case where $h(x, u, \nabla u) := \sigma|\nabla u|$, $\sigma > 0$, we obtain

$$\begin{aligned} H(u,\Omega) = &\int_\Omega \big\{ \sigma\,|\nabla u(x)| + \varphi(x)\cdot \nabla u^T(x) \nabla_u \theta(x,u(x)) \big\}\, dx \\ &+ \int_\Omega \theta(x,u(x)) \operatorname{div}\varphi(x)\, dx + \int_\Omega \varphi(x)\cdot \nabla_x \theta(x,u(x))\, dx, \end{aligned}$$

and it is clear that the functional H is not bounded from below in general, unless one assumes a condition of the type

$$|\nabla_u \theta(x,u)| \le \sigma \qquad \text{for a.e. } x \in \Omega \quad \text{and for all } u \in \mathbb{R}^p,$$

which is essentially the condition found by Massari and Pepe [MP] and by Modica [Mo2]. We note that once (1.1) is written in the form (1.6), well-known integral

representation results for relaxation of bulk energy functionals apply (see [FM2, FR]). This is illustrated first in section 2 in the scalar case $p = 1$ and for a class of energies including the typical ones considered by Visintin (see [V1, V2]),

$$h(x, u, \xi) := \sigma\,|\xi|, \qquad \theta(x, u) := \hat{\sigma}\,u.$$

More general models, relevant to the study of anisotropic materials and nonlinear contact forces, are studied in section 3. In particular, degenerate bounds for the bulk energy density h, as introduced in [FM2] in the vectorial case, may prove to be useful in the analysis of phase transition problems.

In the second part of the paper, sections 4 and 5, we are concerned with constrained minimization problems of the type

$$\inf\left\{\mathcal{H}(u,\Omega) + \int_\Omega \tau(x, u(x))\,dx :\ u \in BV(\Omega; K)\right\},$$

where K is a nonempty compact set of $\mathbb{R}^p$, and $\tau : \Omega \times K \to \mathbb{R}$ is a continuous function. These kinds of problems have important applications in the study of phase transformations and in nucleation phenomena (cf. [V1, V2]). According to the van der Waals–Cahn–Hilliard theory of phase transitions (cf. [CH1, CH2, vdW]), the total energy of a fluid of total mass m and density $u(x)$, confined in a bounded container $\Omega \subset \mathbb{R}^N$, is given by

(1.7)
$$E_\varepsilon(u) := \varepsilon^2 \int_\Omega |\nabla u|^2\,dx + \int_\Omega W_1(u)\,dx + \varepsilon \int_{\partial\Omega} W_2(T\,u)\,dH_{N-1}, \qquad u \in W^{1,1}(\Omega;\mathbb{R}),$$

where the coarse-grain energy $W_1(u)$ is a double well potential vanishing only at α and β and corresponding to the stable two-phase configuration of the fluid, the gradient term $\varepsilon^2|\nabla u|^2$ models the interfacial energy across a smooth transition layer, with ε a small parameter, and W_2 represents the contact energy between the fluid and the container walls. The stable configurations of the fluid correspond to solutions of the problem (see [C])

$$\inf\left\{E_\varepsilon(u) :\ u \in W^{1,1}(\Omega;\mathbb{R}),\ \int_\Omega u\,dx = m\right\}.$$

Confirming a conjecture of Gurtin [G], Modica in [Mo2] was able to show that if a sequence of minimizers $\{u_\varepsilon\}$ converges in L^1 to a function u_0, then u_0 solves the *liquid-drop problem*

$$\inf\left\{\mathcal{H}(u,\Omega) :\ u \in BV(\Omega;\{\alpha,\,\beta\})\right\},$$

where $\mathcal{H}(u,\Omega)$ is the relaxed energy of

$$H(u,\Omega) = \int_\Omega |\nabla u|\,dx + \hat{\sigma}\int_{\partial\Omega} T\,u\,dH_{N-1}, \qquad u \in W^{1,1}(\Omega;\mathbb{R}).$$

Here $\hat{\sigma}$ depends only on W_1 and W_2. The liquid-drop problem admits a solution if and only if $|\hat{\sigma}| \le 1$. An analogous result is due to Alberti, Bouchitté, and Seppecher [ABS] who recently showed that if the parameter ε in front of the contact energy in (1.7) is replaced by λ_ε, where

$$\lim_{\varepsilon\to 0} \varepsilon \log \lambda_\varepsilon = K \in (0,\infty)$$

and W_2 is a double well potential which vanishes only at α_1 and β_1, then the limit problem is given by a different model for capillarity with line tension. It is worth noting that in this case the effective energy takes the form

$$\mathcal{H}(u,\Omega) = \int_\Omega |D(G(u))| + \inf\left\{\int_{\partial\Omega} |G(T\,u) - G(v)|\,dH_{N-1} + \frac{K}{\pi}\int_{\partial\Omega} |Dv|^2 : v \in BV(\partial\Omega;\{\alpha_1,\beta_1\})\right\}$$

for $u \in BV(\Omega;\{\alpha,\beta\})$ and $\mathcal{H}(u,\Omega) = \infty$ otherwise. Here G is a primitive of $2\sqrt{W_1}$. It can be seen immediately that in this capillarity model the contact energy is nonlocal and strongly nonlinear, and again this leads us to consider functions θ other than $\theta(x,u) = \hat{\sigma}\,u$ (see [V1, V2] and section 3).

In the last section of the paper we prove some minimization results which are related to solid nucleation. For a complete description of this phenomenon we refer to the recent monograph of Visintin [V1] and to the bibliography contained therein. By *solid nucleation* we mean the formation of a new solid phase, that is, of a connected component of solid in a liquid. If the new solid phase is formed in the interior of the liquid, the nucleation is called *homogeneous*, while if it is also in contact with other substances, such as the container, impurities dispersed in the liquid or nucleants, then we name it *heterogeneous* nucleation (cf. [V1, Ch. VII.2]). By thinking of these impurities or particles as holes in the domain Ω, we can represent the contact energy by an integral term over the boundary of Ω. Furthermore, since the new solid phase is formed through crystallization, and crystals are *anisotropic*, the classical isotropic interfacial energy $\sigma\int_\Omega |Du|$ is now replaced by $\int_\Omega h(x,Du)$. In the applications one sees often $h(x,Du) = |A(x)Du|$, where $A(x)$ is a nonnegative definite $N\times N$ tensor (cf. [V1, p. 157]).

The main results of this part are Theorems 5.1 and 5.4, where we show that minimizers u of the total energy are given by pure phases; that is, there exists $\Omega_u \subset \Omega$ such that $u(x) = \alpha$ for $\mathcal{L}^N$ a.e. $x \in \Omega_u$ (liquid) and $u(x) = \beta$ for $\mathcal{L}^N$ a.e. $x \in \Omega\backslash\Omega_u$ (solid). This result is closely related to Theorem 2 in [V2], where the interfacial energy is assumed to satisfy a *generalized co-area formula*. We replace here this condition by some hypotheses which are easy to verify and allow us to include interfacial energies of the form $\int_\Omega h(x,Du)$, where $h(x,\cdot)$ is convex and positively homogeneous of degree one, and for which the co-area formula might not hold.

2. Relaxation: A simple model problem. In what follows $\Omega \subset \mathbb{R}^N$ is an open, bounded set of class C^2, and T stands for the trace operator on $\partial\Omega$.

In this section we characterize the relaxed energy $\mathcal{H}$ when

(A_1) $h : \Omega\times\mathbb{R}^N \to [0,+\infty)$ is a continuous function;

(A_2) $h(x,\cdot)$ is convex for all $x \in \Omega$;

(A_3) $\sigma|\xi| \le h(x,\xi) \le C(1+|\xi|)$, for some $\sigma, C > 0$, and for all $(x,\xi) \in \Omega\times\times\mathbb{R}^N$;

(A_4) $\theta(x,u) := \hat{\sigma}\,u$, for some $|\hat{\sigma}| \le \sigma$ and for all $(x,u) \in \Omega\times\mathbb{R}$.

The main idea is to rewrite

$$H(u,\Omega) := \int_\Omega h(x,\nabla u(x))\,dx + \int_{\partial\Omega} \hat{\sigma}\,T\,u(x)\,dH_{N-1}(x), \quad u \in W^{1,1}(\Omega;\mathbb{R}^p),$$

as a bulk energy and then apply known relaxation results from $W^{1,1}$ to BV (see, e.g., [DM, FM2]).

We start with some notation. For any $\nu \in S^{N-1} := \{x \in \mathbb{R}^N : |x| = 1\}$ let $\{\nu_1,\dots,\nu_{N-1},\nu\}$ be an orthonormal basis of $\mathbb{R}^N$ varying continuously with ν, and

$Q_\nu := \{x \in \mathbb{R}^N : |x\cdot\nu_i| < 1/2,\ |x\cdot\nu| < 1/2,\ i = 1,\dots,N-1\}$ be a unit cube centered at the origin with two of its faces orthogonal to the direction ν.

If g is a positively homogeneous function of degree one and if μ is an $\mathbb{R}^m$-valued measure, then we define

$$\int_\Omega g(d\mu) := \int_\Omega g(\alpha(x))\, d|\mu(x)|,$$

where $|\mu|$ is the nonnegative total variation measure of μ, and $\alpha : \Omega \to S^{m-1}$ is the Radon–Nikodym derivative of μ with respect to $|\mu|$.

We recall briefly some facts about functions of bounded variation which will be useful in the sequel. A function $u \in L^1(\Omega;\mathbb{R}^p)$ is said to be of *bounded variation* if for all $i = 1,\dots,p$, and $j = 1,\dots,N$, there exists a Radon measure μ_{ij} such that

$$\int_\Omega u_i(x)\,\frac{\partial\varphi}{\partial x_j}\,(x)\,dx = -\int_\Omega \varphi(x)\,d\mu_{ij}$$

for every $\varphi \in C_0^1(\Omega;\mathbb{R})$. The distributional derivative Du is the matrix-valued measure with components μ_{ij}. Given $u \in BV(\Omega;\mathbb{R}^p)$ the *approximate upper* and *lower limit* of each component u_i, $i = 1,\dots,p$, are given by

$$u_i^+(x) := \inf\left\{t \in \mathbb{R} : \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon^N}\,\mathcal{L}^N\left(\{y \in \Omega \cap B(x,\varepsilon) :\ u_i(y) > t\}\right) = 0\right\}$$

and

$$u_i^-(x) := \sup\left\{t \in \mathbb{R} : \lim_{\varepsilon\to 0^+} \frac{1}{\varepsilon^N}\,\mathcal{L}^N\left(\{y \in \Omega \cap B(x,\varepsilon) :\ u_i(y) < t\}\right) = 0\right\},$$

while the *jump set* of u, or *singular set*, is defined by

$$S(u) := \cup_{i=1}^p \{x \in \Omega :\ u_i^-(x) < u_i^+(x)\}.$$

It is well known that $S(u)$ is $N-1$ rectifiable; i.e.,

$$S(u) = \cup_{n=1}^\infty K_n \cup E,$$

where $H_{N-1}(E) = 0$ and K_n is a compact subset of a C^1 hypersurface. If $x \in \Omega\backslash S(u)$, then $u(x)$ is taken as the common value of $(u_1^+(x),\dots,u_p^+(x))$ and $(u_1^-(x),\dots,u_p^-(x))$. It can be shown that $u(x) \in \mathbb{R}^p$ for H_{N-1} a.e. $x \in \Omega\backslash S(u)$. Furthermore, for H_{N-1} a.e. $x \in S(u)$ there exist a unit vector $\nu_u(x) \in S^{N-1}$, normal to $S(u)$ at x, and two vectors $u^-(x)$, $u^+(x) \in \mathbb{R}^p$ (the traces of u on $S(u)$ at the point x) such that

$$\lim_{r\to 0} \frac{1}{r^N} \int_{\{y\in B(x_0,r):\,(y-x)\cdot\nu_u(x)>0\}} |u(y) - u^+(x)|^{N/(N-1)} dy = 0$$

and

$$\lim_{r\to 0} \frac{1}{r^N} \int_{\{y\in B(x_0,r):\,(y-x)\cdot\nu_u(x)<0\}} |u(y) - u^-(x)|^{N/(N-1)} dy = 0.$$

Note that in general $(u_i)^+ \neq (u^+)_i$ and $(u_i)^- \neq (u^-)_i$. Moreover, the Sobolev inequality

$$\left(\int_\Omega |u(x)|^{N/(N-1)} dx\right)^{(N-1)/N} \leq C(N)||u||_{BV}$$

holds in $BV(\Omega;\mathbb{R}^p)$ when $N>1$. Finally, Du may be represented as

$$Du = \nabla u\,\mathcal{L}^N + (u^+ - u^-)\otimes \nu_u\, H_{N-1}\lfloor S(u) + C(u),$$

where ∇u is the density of the absolutely continuous part of Du with respect to the N-dimensional Lebesgue measure $\mathcal{L}^N$. These three measures are mutually singular.

LEMMA 2.1. *There exists $\varphi\in C_0^1(\mathbb{R}^N;\mathbb{R}^N)$ with $|\varphi(x)|<1$ in Ω such that for any $v\in BV(\Omega;\mathbb{R}^p)$,*

$$\int_{\partial\Omega} T\,v(x)\,dH_{N-1}(x) = \int_\Omega v(x)\mathrm{div}\,\varphi(x)\,dx + \int_\Omega \varphi(x)\cdot d(D(v(x))).$$

Proof. Since $\partial\Omega$ is compact and of class C^2, we can find a finite open covering $\{U_j\}_j$ of $\partial\Omega$, where U_j are balls centered at points of $\partial\Omega$, $j=1,\dots,P$, and for each U_j there is a C^2 diffeomorphism $\Phi_j: U_j\to\Phi_j(U_j)$ such that $\Phi_j(U_j)\subset B(0,R_j)\subset\mathbb{R}^N$ for some $R_j>0$,

$$\Omega\cap U_j = \{x\in U_j:\,(\Phi_j(x))_N<0\} \tag{2.1}$$

and for $x\in\partial\Omega\cap U_j$ the exterior normal to $\partial\Omega$ at x is given by

$$n(x,\Omega) = \frac{\nabla\Phi_j^T(x)\,e_N}{|\nabla\Phi_j^T(x)\,e_N|}.$$

Let Ψ be a partition of the unity for $\cup_{j=1}^P U_j$ subordinate to $\{U_j\}_j$. For any $\psi\in\Psi$ there exists $j\in\{1,\dots,P\}$ such that $\psi\in C_0^\infty(U_j)$, and we define

$$\varphi_\psi(x) := \frac{\nabla\Phi_j^T(x)\,e_N}{|\nabla\Phi_j^T(x)\,e_N|}\left(1+\frac{(\Phi_j(x))_N}{R_j}\right)\psi(x); \tag{2.2}$$

then $\varphi_\psi(x)\in C_0^1(U_j;\mathbb{R}^N)$ and $|\varphi_\psi(x)|<1$ for $x\in\Omega\cap U_j$. If we set φ_ψ to be zero outside U_j we obtain that $\varphi_\psi(x)\in C_0^1(\mathbb{R}^N;\mathbb{R}^N)$, and thus we can apply the trace theorem (cf. [EG, Thm. 5.3.1]) to the BV function v, and we have

$$\begin{aligned}\int_{\partial\Omega}\varphi_\psi(x)\cdot n(x,\Omega)\,T\,v(x)\,dH_{N-1}(x)&\\ = \int_\Omega v(x)\mathrm{div}\,\varphi_\psi(x)\,dx + \int_\Omega \varphi_\psi(x)\cdot d(D(v(x))).&\end{aligned}$$

On the other hand, since by (2.1) $\varphi_\psi(x) = n(x,\Omega)\,\psi(x)$ if $x\in\partial\Omega\cap U_j$, while $\varphi_\psi(x)=0$ if $x\in\partial\Omega\backslash U_j$, we get

$$\int_{\partial\Omega}\varphi_\psi(x)\cdot n(x,\Omega)\,T\,v(x)\,dH_{N-1}(x) = \int_{\partial\Omega}\psi(x)T\,v(x)\,dH_{N-1}(x).$$

Hence

$$\begin{aligned}\int_{\partial\Omega} T\,v(x)\,dH_{N-1}(x) &= \sum_{\psi\in\Psi}\int_{\partial\Omega}\psi(x)T\,v(x)\,dH_{N-1}(x)\\ &= \int_\Omega v(x)\,\mathrm{div}\left(\sum_{\psi\in\Psi}\varphi_\psi(x)\right)dx + \int_\Omega\left(\sum_{\psi\in\Psi}\varphi_\psi(x)\right)\cdot d(D(v(x))).\end{aligned}$$

The proof of Lemma 2.1 is complete if we show that $\varphi(x) := \sum_{\psi\in\Psi} \varphi_\psi(x)$ satisfies $|\varphi(x)| < 1$ in Ω. Fix $x \in \Omega$. If $x \notin \cup_{j=1}^P U_j$, then $\varphi(x) = 0$. If $x \in \cup_{j=1}^P U_j$, then $\sum_{\psi\in\Psi} \psi(x) = 1$, and so there exists at least one $\psi_0 \in \Psi$ such that $\psi_0(x) > 0$. Let $j \in \{1, \dots, P\}$ be such that $\psi_0 \in C_0^\infty(U_j)$. Then by (2.1) and (2.2)

$$|\varphi_{\psi_0}(x)| = \left(1 + \frac{(\Phi_j(x))_N}{R_j}\right) \psi_0(x) < \psi_0(x),$$

and consequently, since $\psi(x) = 0$ for all but finitely many $\psi \in \Psi$,

$$|\varphi(x)| \le \sum_{\psi\in\Psi} |\varphi_\psi(x)| < \sum_{\psi\in\Psi} \psi(x) = 1. \qquad \square$$

Using this lemma, we are now in position to obtain an integral representation for the relaxed energy

$$\mathcal{H}(u,\Omega) := \inf_{\{u_n\}} \left\{ \liminf_{n\to\infty} H(u_n,\Omega) :\ u_n \in W^{1,1}(\Omega;\mathbb{R}^p),\ u_n \to u \quad \text{in} \quad L^1(\Omega;\mathbb{R}^p) \right\}.$$

THEOREM 2.2. *We have that*

$$\begin{aligned}\mathcal{H}(u,\Omega) = &\int_\Omega h(x,\nabla u(x))\,dx + \int_\Omega h^\infty(x, dC(u)) \\ &+ \int_{S(u)\cap\Omega} h^\infty(x, (u^+(x) - u^+(x)) \otimes \nu_u(x))\, dH_{N-1}(x) \\ &+ \int_{\partial\Omega} \hat\sigma\, T\, u(x)\, dH_{N-1}(x),\end{aligned}$$

where the recession function h^∞ of h is defined as

$$h^\infty(x,\xi) := \lim_{t\to\infty} \frac{h(x, t\,\xi)}{t}.$$

Proof. In light of Lemma 2.1 we may rewrite the energy $H(u,\Omega)$ as

$$\begin{aligned}H(u,\Omega) &= \int_\Omega [h(x,\nabla u(x)) + \hat\sigma\, \varphi(x)\cdot\nabla u(x)]\,dx + \int_\Omega \hat\sigma\, u(x)\mathrm{div}\varphi(x)\,dx \\ &= \int_\Omega f(x,\nabla u(x))\,dx + \int_\Omega \hat\sigma\, u(x)\mathrm{div}\varphi(x)\,dx,\end{aligned}$$

where $|\varphi| < 1$ in Ω, and $f(x,\xi) := h(x,\xi) + \hat\sigma\,\varphi(x)\cdot\xi$. Since

$$(\sigma - \hat\sigma\,|\varphi(x)|)|\xi| \le f(x,\xi) \le C'(1 + |\xi|)$$

for some $C' > 0$, we may use the relaxation arguments introduced in [FM1, FM2], taking into account that the relaxation method is *local*, so that on each ball $B(x_0,\varepsilon) \subset\subset \Omega$ there is an upper bound $|\varphi(x)| < \alpha < 1$ and thus f is coercive. Therefore, and in view of the strong continuity in L^1 of the mapping

$$u \mapsto \int_\Omega \hat\sigma\, u(x)\mathrm{div}\varphi(x)\,dx,$$

we conclude that (see also [ADM2, DM])

$$\begin{aligned}\mathcal{H}(u,\Omega) &= \int_\Omega f(x,\nabla u(x))\,dx + \int_\Omega f^\infty(x,dC(u))\\ &+ \int_{S(u)\cap\Omega} f^\infty(x,(u^+(x)-u^+(x))\otimes\nu_u(x))\,dH_{N-1}(x) + \int_\Omega \hat\sigma\, u(x)\mathrm{div}\varphi(x)\,dx\\ &= \int_\Omega h(x,\nabla u(x))\,dx + \int_\Omega h^\infty(x,dC(u))\\ &+ \int_{S(u)\cap\Omega} h^\infty(x,(u^+(x)-u^+(x))\otimes\nu_u(x))\,dH_{N-1}(x) + \int_{\partial\Omega}\hat\sigma\, T\,u(x)\,dH_{N-1}(x),\end{aligned}$$

where we have used once again Lemma 2.1 in the last equality. □

3. Relaxation: A general model. Here we consider a more general model corresponding to the functional

$$H(u,\Omega) := \int_\Omega h(x,u(x),\nabla u(x))\,dx + \int_{\partial\Omega}\theta(x,T\,u(x))\,dH_{N-1}(x)$$

defined on the Sobolev space $W^{1,1}(\Omega;\mathbb{R}^p)$, where $\Omega\subset\mathbb{R}^N$ is an open, bounded set of class C^2, T is the trace operator on $\partial\Omega$, H_{N-1} is the $N-1$-dimensional Hausdorff measure, and the functions

$$h:\Omega\times\mathbb{R}^p\times\mathbb{M}^{p\times N}\to[0,\infty),\qquad \theta:\partial\Omega\times\mathbb{R}^p\to\mathbb{R}$$

satisfy the following hypotheses:

(H_1) h is continuous;

(H_2) $h(x,u,\cdot)$ is quasiconvex for all $(x,u)\in\Omega\times\mathbb{R}^p$;

(H_3) there exist a nonnegative, bounded, continuous function $g:\Omega\times\mathbb{R}^p\to[0,\infty)$ and a constant $C>0$ such that

$$g(x,u)|\xi| \le h(x,u,\xi) \le C\,g(x,u)(1+|\xi|) \tag{3.1}$$

for all $(x,u,\xi)\in\Omega\times\mathbb{R}^p\times\mathbb{M}^{p\times N}$, where $\mathbb{M}^{p\times N}$ is the vector space of $p\times N$ matrices;

(H_4) for every $x_0\in\Omega$ and $\delta>0$ there exists $\varepsilon>0$ such that

$$h(x_0,u,\xi) - h(x,u,\xi) \le \delta(1+g(x,u)\,|\xi|) \tag{3.2}$$

for all $x\in\Omega$ with $|x-x_0|\le\varepsilon$ and for all $(u,\xi)\in\mathbb{R}^p\times\mathbb{M}^{p\times N}$;

(H_5) there exist $C'>0$ and $m\in(0,1)$ such that

$$|h^\infty(x,u,\xi) - h(x,u,\xi)| \le C'g(x,u)(1+|\xi|^{1-m})$$

for all $(x,u,\xi)\in\Omega\times\mathbb{R}^p\times\mathbb{M}^{p\times N}$, where the *recession function* h^∞ of h is defined as

$$h^\infty(x,u,\xi) := \limsup_{t\to\infty}\frac{h(x,u,t\,\xi)}{t};$$

(H_6) θ admits an extension $\theta\in C(\overline{\Omega}\times\mathbb{R}^p;\mathbb{R})\cap C^1(\Omega\times\mathbb{R}^p;\mathbb{R})$ such that

$$|\nabla_x\theta(x,u)| \le a_1(x) + C_1(1+|u|^{q_c})$$

for $\mathcal{L}^N$ a.e. $x \in \Omega$ and all $u \in \mathbb{R}^p$, where $a_1 \in L^1(\Omega, \mathbb{R})$, $C_1 > 0$, and q_c is the Sobolev exponent $q_c := N/(N-1)$ if $N > 1$ and $q_c < \infty$ if $N = 1$. Moreover, for every $x_0 \in \Omega$ and $\delta > 0$ there exists $\varepsilon > 0$ such that

$$|\nabla_u \theta(x_0, u) - \nabla_u \theta(x, u)| \leq \delta\, g(x, u) \tag{3.3}$$

for all $x \in \Omega$ with $|x - x_0| \leq \varepsilon$ and for all $u \in \mathbb{R}^p$;

(H_7) $g(x, u) \geq |\nabla_u \theta(x, u)|$ for all $(x, u) \in \Omega \times \mathbb{R}^p$.

Remark 3.1. (i) Conditions (H_1)–(H_5) were considered by Fonseca and Müller (see [FM2]), who treated the case where $\theta \equiv 0$. Actually, in [FM2] hypothesis (H4) included:

(H4)$_1$ For every compact set $K \Subset \Omega \times \mathbb{R}^p$ there exists a continuous function $\omega : [0, \infty) \to [0, \infty)$, with $\omega(0) = 0$, such that

$$|h(x, u, \xi) - h(x_1, u_1, \xi)| \leq \omega(|x - x_1| + |u - u_1|)(1 + |\xi|)$$

for all (x, u, ξ), $(x_1, u_1, \xi) \in K \times \mathbb{M}^{p \times N}$.

As it turns out, this property follows from (H1), (H2), (H5). Indeed using the fact that the recession function h^∞ of h is still quasiconvex and is positively homogeneous of degree one in the ξ variable (see [FM2, M]), by (H1), (H2), (H5), it is possible to show that h^∞ is actually continuous in $\Omega \times \mathbb{R}^p \times \mathbb{M}^{p \times N}$ and, in turn, that h satisfies condition (H4)$_1$ of [FM2]. We omit the details.

(ii) By the mean value theorem and conditions (H_3) and (H_7) we have

$$|\theta(x, u) - \theta(x, u_1)| \leq |\nabla_u \theta(x, \hat{u})|\, |u - u_1| \leq ||g||_{L^\infty} |u - u_1| \tag{3.4}$$

for all $x \in \Omega$ and all u, $u_1 \in \mathbb{R}^p$. Taking $u_1 = 0$ it follows by (H_6) and (3.4) that

$$|\theta(x, u)| \leq ||g||_{L^\infty}\, |u| + ||\theta(x, 0)||_{L^\infty} \tag{3.5}$$

for all $(x, u) \in \Omega \times \mathbb{R}^p$. This growth condition, together with (3.1), implies in particular that the functional $H(u, \Omega)$ is well defined and finite for $u \in W^{1,1}(\Omega; \mathbb{R}^p)$.

(iii) A typical example of the energy densities is (see Visintin [V1, V2] and section 2)

$$h(x, u, \xi) := \sigma\, |\xi|, \qquad \theta(x, u) := \hat{\sigma}\, u, \tag{3.6}$$

where $\sigma > 0$ and $\hat{\sigma} \in \mathbb{R}$. It is easy to see that conditions (H_1)–(H_6) hold with $g(x, u) := \sigma$, while assumption (H_7) reduces to the inequality $|\hat{\sigma}| \leq \sigma$. More generally, ($H_6$) is trivially satisfied if $\theta = \theta(u)$.

(iv) If in (1.3) we take $\theta(x, u) := \hat{\sigma}|u|$ for $(x, u) \in \partial\Omega \times \mathbb{R}^p$ (cf. [MP]), then it is possible to extend θ to $\overline{\Omega} \times \mathbb{R}^p$ as follows:

$$\theta(x, u) := \hat{\sigma} \sqrt{|u|^2 + \psi^2(x)},$$

where $\psi \in C^1(\overline{\Omega}; \mathbb{R})$ is such that $\psi(x) > 0$ for $x \in \Omega$ and $\psi(x) = 0$ for $x \in \partial\Omega$. Conditions (H_1)–(H_7) are then verified with $g(x, u) := \sigma$, provided $|\hat{\sigma}| \leq \sigma$. The problem of finding an extension of $\theta : \partial\Omega \times \mathbb{R}^p \to \mathbb{R}$ to $\overline{\Omega} \times \mathbb{R}^p$ which satisfies (H_6)–(H_7) for the functional (1.3), and when (1.4) holds, will be addressed in a forthcoming paper.

Our goal in this section is to generalize the integral representation obtained in section 2 to the relaxed energy of $H(u, \Omega)$ in $BV(\Omega; \mathbb{R}^p)$ with respect to the L^1 topology; that is,

$$\mathcal{H}(u, \Omega) := \inf_{\{u_n\}} \left\{ \liminf_{n \to \infty} H(u_n, \Omega) :\ u_n \in W^{1,1}(\Omega; \mathbb{R}^p),\ u_n \to u \quad \text{in} \quad L^1(\Omega; \mathbb{R}^p) \right\}.$$

From the definition of $\mathcal{H}(u,\Omega)$ it follows immediately that the functional $\mathcal{H}(u,\Omega)$ is lower semicontinuous in $L^1(\Omega;\mathbb{R}^p)$.

Before stating the main theorems of this section we introduce the *surface energy* associated with the function h. For fixed a, $b \in \mathbb{R}^p$ we define $\mathcal{A}(a,b,\nu)$ as the class of all functions $\psi \in W^{1,1}(Q_\nu;\mathbb{R}^p)$ such that

$$T\,\psi(y) = \begin{cases} a & \text{if } y\cdot\nu = -1/2, \\ b & \text{if } y\cdot\nu = \ \ 1/2, \end{cases}$$

and which are periodic of period one in the remaining directions $\nu_1,\dots,\nu_{N-1}$. The surface energy associated with the function h, $K_h(x,a,b,\nu)$, is defined by

$$K_h(x,a,b,\nu) := \inf\left\{\int_{Q_\nu} h^\infty(x,\psi(y),\nabla\psi(y))\,dy :\ \psi\in\mathcal{A}(a,b,\nu)\right\}.$$

For a detailed study of the properties of the function $K_h(x,a,b,\nu)$ we refer to [FR].

For $u \in BV(\Omega;\mathbb{R}^p)$ we define the functional

$$\begin{aligned}\mathcal{L}(u,\Omega) := &\int_\Omega h(x,u(x),\nabla u(x))\,dx + \int_\Omega h^\infty(x,u(x),dC(u)) \\ &+ \int_{S(u)\cap\Omega} K_h(x,u^-(x),u^+(x),\nu_u(x))\,dH_{N-1}(x) + \int_{\partial\Omega}\theta(x,T\,u(x))\,dH_{N-1}(x).\end{aligned}$$

THEOREM 3.2. *Let* (H_1)–(H_7) *hold. If* $u\in BV(\Omega;\mathbb{R}^p)$, *then*

$$\mathcal{H}(u,\Omega) = \mathcal{L}(u,\Omega).$$

COROLLARY 3.3. *If* $h = h(x,\xi)$, *then*

$$\begin{aligned}\mathcal{H}(u,\Omega) = &\int_\Omega h(x,\nabla u(x))\,dx + \int_\Omega h^\infty(x,dC(u)) \\ &+ \int_{S(u)\cap\Omega} h^\infty(x,(u^+(x)-u^-(x))\otimes\nu_u(x))\,dH_{N-1}(x) \\ &+ \int_{\partial\Omega}\theta(x,T\,u(x))\,dH_{N-1}(x).\end{aligned}$$

The proof of Corollary 3.3 follows from Remark 2.17 in [FM2].

Remark 3.4. (i) Rather surprisingly, in general the functional $\mathcal{L}(u,\Omega)$ is not lower semicontinuous in L^1 if the domain Ω is only Lipschitz. This fact was first pointed out by Modica in [Mo2] who gave the following simple example. Let $\Omega := (0,1)\times(0,1)\subset\mathbb{R}^2$ and take h and θ as in (3.6), with $-\sigma \le \hat\sigma < -\sigma\sqrt{2}/2$. Then (H_1)–(H_7) are satisfied (see Remark 3.1(iii)), and

$$\mathcal{L}(u,\Omega) = \sigma\int_\Omega |Du| + \hat\sigma\int_{\partial\Omega} T\,u\,dH_1, \quad u\in BV(\Omega;\mathbb{R}).$$

Consider the sequence

$$u_n(x_1,x_2) := \begin{cases} 0 & \text{if } x_1+x_2 \ge 1/n, \\ n & \text{if } x_1+x_2 < 1/n. \end{cases}$$

Then $u_n(x)\to 0$ in $L^1(\Omega;\mathbb{R})$ but $\mathcal{L}(u_n,\Omega) = \sigma\sqrt{2}+2\hat\sigma < \mathcal{L}(0,\Omega) = 0$, and this shows that $\mathcal{H}(u,\Omega)\ne\mathcal{L}(u,\Omega)$ since $\mathcal{H}(u,\Omega)$ is lower semicontinuous in L^1.

It is worth noting that in the special case where $\theta(x,u) = \hat{\sigma}\,|u-\psi(x)|$ in (1.3), with $|\hat{\sigma}| \le \sigma$ and $\psi \in L^1(\partial\Omega;\mathbb{R})$, one can still prove lower semicontinuity of $\mathcal{L}$ for Lipschitz subdomains of $\mathbb{R}^N$. The first result in this direction is due to Massari and Pepe [MP] who treated the case where $\psi \equiv 0$. Modica [Mo2] then extended it to include $\psi \in L^1(\partial\Omega;\mathbb{R})$. The idea in [MP, Mo2] is to find a function $\hat{\psi} \in BV(\mathbb{R}^N \setminus \overline{\Omega};\mathbb{R})$ whose trace is ψ and then use an extension theorem (see [EG, Thm. 5.4.1]) to rewrite the integral $\int_{\partial\Omega} |T\,u - \psi|\,dH_{N-1}$ as

$$\int_{\partial\Omega} |T\,u - T\,\hat{\psi}|\,dH_{N-1} = \int_{\mathbb{R}^N} |D\hat{u}| - \int_{\Omega} |Du| - \int_{\mathbb{R}^N \setminus \overline{\Omega}} |D\hat{\psi}|,$$

where

$$\hat{u}(x) := \begin{cases} u(x) & \text{if } x \in \Omega, \\ \hat{\psi}(x) & \text{if } x \in \mathbb{R}^N \setminus \overline{\Omega}. \end{cases}$$

(ii) Without condition (H_7) Theorem 3.2 may fail. As an example, let $\Omega := (0,1) \subset \mathbb{R}$ and take h and θ as in (3.6). In this case condition (H_7) is equivalent to the inequality $|\hat{\sigma}| \le \sigma$. Assume that $\sigma < \hat{\sigma}$ and consider the sequence

$$u_n(x) := \begin{cases} -n^3(x-1) - n & \text{if } 1 - 1/n^2 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then $u_n(x) \to 0$ in $L^1(\Omega;\mathbb{R})$ but $\mathcal{L}(u_n,\Omega) = (\sigma - \hat{\sigma})n < \mathcal{L}(0,\Omega) = 0$.

THEOREM 3.5. *Let* (H_1)–(H_6) *hold, with* (3.1) *and* (H_7) *replaced by the weaker hypothesis*

$$|\nabla_u \theta(x,u)|\,|\xi| \le h(x,u,\xi) \le C\,g(x,u)(1+|\xi|) \tag{3.7}$$

for all $(x,u,\xi) \in \Omega \times \mathbb{R}^p \times \mathbb{M}^{p\times N}$, *and some* $C > 0$. *Then the relaxation of* $H(u,\Omega)$

$$\mathcal{H}_b(u,\Omega) = \inf_{\{u_n\}} \Big\{ \liminf_{n\to\infty} H(u_n,\Omega):\ u_n \in W^{1,1}(\Omega;\mathbb{R}^p),\ u_n \to u \text{ in } L^1(\Omega;\mathbb{R}^p),\ \sup_n ||u_n||_{W^{1,1}} < \infty \Big\}$$

in $BV(\Omega;\mathbb{R}^p)$ *with respect to the* L^1 *topology has the integral representation*

$$\mathcal{H}_b(u,\Omega) = \mathcal{L}(u,\Omega).$$

Remark 3.6. Under the assumptions of Theorem 3.5, the functional $\mathcal{L}(u,\Omega)$ provides the correct integral representation for $\mathcal{H}_b(u,\Omega)$ but not necessarily for $\mathcal{H}(u,\Omega)$. Indeed, in the scalar case where $p = 1$ and when $\theta \equiv 0$, Dal Maso has shown in [DM] that $\mathcal{H}(u,\Omega) = \mathcal{L}(u,\Omega)$ when $h = h(u,\xi)$ satisfies only (3.7), while possibly $\mathcal{H}(u,\Omega) < \mathcal{L}(u,\Omega)$ for $h = h(x,\xi)$ unless one assumes a condition of the type (3.1).

In the vectorial case where $p > 1$ and when $\theta \equiv 0$, Ambrosio and Dal Maso [ADM2] proved that $\mathcal{H}(u,\Omega) = \mathcal{L}(u,\Omega)$ when $h = h(\xi)$ satisfies only (3.7). Independently, Fonseca and Müller [FM2] have obtained this result for general functions $h(x,u,\xi)$ which verify (3.1), still in the case where $\theta \equiv 0$.

We proceed with the proofs of Theorems 3.2 and 3.5. We start with some preliminary results. In what follows, and unless otherwise specified, we always assume that conditions (H_1)–(H_7) hold.

LEMMA 3.7. *If $u \in BV(\Omega;\mathbb{R}^p)$, then the function $v(x) := \theta(x, u(x)) \in BV(\Omega;\mathbb{R})$ and*

$$Dv = \begin{cases} \nabla_x \theta(x,u)\, \mathcal{L}^N + Du^T \nabla_u \theta(x,u) & \text{on } \Omega\backslash S(u), \\ (\theta(x,u^+) - \theta(x,u^-)) \otimes \nu_u \, H_{N-1} \lfloor S(u) & \text{on } S(u). \end{cases}$$

Moreover

$$T\, v(x) = \theta(x, T\, u(x)).$$

The proof of Lemma 3.7 is straightforward in light of related results on the chain rule for BV functions (see [ADM1] and the references contained therein).

By Lemmas 2.1 and 3.7, if $u \in BV(\Omega;\mathbb{R}^p)$, then

(3.8)
$$\int_{\partial\Omega} \theta(x, T\, u(x))\, dH_{N-1}(x) = \int_\Omega \theta(x,u(x)) \mathrm{div}\, \varphi(x)\, dx + \int_\Omega \varphi(x) \cdot d(D(\theta(x,u(x)))).$$

In turn, by (3.8) we can rewrite the functional $H(u,\Omega)$ as

(3.9)
$$\begin{aligned} H(u,\Omega) = &\int_\Omega \left\{ h(x,u(x),\nabla u(x)) + \varphi(x) \cdot \nabla u^T(x) \nabla_u \theta(x,u(x)) \right\} dx \\ &+ \int_\Omega \theta(x,u(x))\, \mathrm{div}\, \varphi(x)\, dx + \int_\Omega \varphi(x) \cdot \nabla_x \theta(x,u(x))\, dx. \end{aligned}$$

This equivalent form gives us a better insight into the competing roles played by the two energy integrals $\int_\Omega h(x,u,\nabla u)\, dx$ and $\int_{\partial\Omega} \theta(x, T\, u)\, dH_{N-1}$. In particular, it is now clear that without a condition of the type

$$h(x,u,\xi) \geq |\nabla_u \theta(x,u)|\, |\xi|$$

one may have $\mathcal{H}(u,\Omega) = -\infty$, as in the example in Remark 3.4(ii).

Define $f(x,u,\xi) := h(x,u,\xi) + \varphi(x) \cdot \xi^T \nabla_u \theta(x,u)$ for $(x,u,\xi) \in \Omega \times \mathbb{R}^p \times \mathbb{M}^{p\times N}$, set

$$F(u,\Omega) := \int_\Omega f(x,u(x),\nabla u(x))\, dx, \quad u \in W^{1,1}(\Omega;\mathbb{R}^p),$$

and let

$$\mathcal{F}(u,\Omega) := \inf_{\{u_n\}} \left\{ \liminf_{n\to\infty} F(u_n,\Omega) :\, u_n \in W^{1,1}(\Omega;\mathbb{R}^p),\, u_n \to u \quad \text{in} \quad L^1(\Omega;\mathbb{R}^p) \right\}.$$

LEMMA 3.8. *If $u \in BV(\Omega;\mathbb{R}^p)$, then*

$$\mathcal{H}(u,\Omega) = \mathcal{F}(u,\Omega) + \int_\Omega \theta(x,u(x))\, \mathrm{div}\, \varphi(x)\, dx + \int_\Omega \varphi(x) \cdot \nabla_x \theta(x,u(x))\, dx.$$

Proof. Clearly it is enough to show that

$$\begin{aligned} \liminf_{n\to\infty} H(u_n,\Omega) = &\liminf_{n\to\infty} F(u_n,\Omega) \\ &+ \int_\Omega \theta(x,u(x))\, \mathrm{div}\, \varphi(x)\, dx + \int_\Omega \varphi(x) \cdot \nabla_x \theta(x,u(x))\, dx \end{aligned}$$

for any sequence $\{u_n\} \subset W^{1,1}(\Omega;\mathbb{R}^p)$ such that $u_n \to u$ in $L^1(\Omega;\mathbb{R}^p)$. We first observe that, since $\varphi \in C_0^1(\mathbb{R}^N;\mathbb{R}^N)$, the functions φ and $\operatorname{div}\varphi$ are bounded in Ω. Moreover, by (3.4)

$$|\theta(x,u_n(x)) - \theta(x,u(x))| \le ||g||_{L^\infty}\,|u_n(x)-u(x)| \quad \text{for } \mathcal{L}^N \text{ a.e. } x \in \Omega.$$

Hence

$$\lim_{n\to\infty}\int_\Omega \theta(x,u_n)\operatorname{div}\varphi\,dx = \int_\Omega \theta(x,u)\operatorname{div}\varphi\,dx.$$

By (H_6), by virtue of the Sobolev inequality, and due to the fact that φ is bounded, the functional $u \mapsto \int_\Omega \varphi\cdot\nabla_x\theta(x,u(x))\,dx$ is continuous in $L^1(\Omega;\mathbb{R}^p)$ (see [K, Thm. 2.1]) and thus

$$\lim_{n\to\infty}\int_\Omega \varphi\cdot\nabla_x\theta(x,u_n)\,dx = \int_\Omega \varphi(x)\cdot\nabla_x\theta(x,u)\,dx. \qquad \square$$

We are now ready to prove Theorem 3.2.

Proof of Theorem 3.2. By Lemma 3.8, in order to find an integral representation for $\mathcal{H}(u,\Omega)$ in $BV(\Omega;\mathbb{R}^p)$ it is sufficient to determine one for $\mathcal{F}(u,\Omega)$. The idea is to apply Theorem 2.16 of [FM2]. In order to do so we need to show that the function

$$f(x,u,\xi) = h(x,u,\xi) + \varphi(x)\cdot\xi^T\nabla_u\theta(x,u)$$

satisfies conditions (H_1)–(H_5) which are essentially the same as [FM2].

Condition (H_1) is trivially verified since the functions θ and φ are of class C^1. As f is the sum of a quasiconvex function and a function linear in ξ, it is clear that $f(x,u,\cdot)$ is still quasiconvex and that

$$f^\infty(x,u,\xi) = h^\infty(x,u,\xi) + \varphi(x)\cdot\xi^T\nabla_u\theta(x,u),$$

which, in turn, implies that

$$|f^\infty(x,u,\xi) - f(x,u,\xi)| = |h^\infty(x,u,\xi) - h(x,u,\xi)| \le C'g(x,u)(1+|\xi|^{1-m})$$

by (H_5). Thus f verifies also (H_2) and (H_5).

We prove (3.2). Fix $x_0 \in \Omega$ and $\delta > 0$. There exists $\varepsilon > 0$ such that for $x \in \Omega$ with $|x-x_0| \le \varepsilon$ and $(u,\xi) \in \mathbb{R}^p \times \mathbb{M}^{p\times N}$,

$$\begin{aligned} h(x_0,u,\xi) - h(x,u,\xi) &\le \tfrac{1}{3}\,\delta(1+g(x,u)|\xi|), \qquad |\varphi(x)-\varphi(x_0)| \le \tfrac{1}{3}\,\delta,\\ |\nabla_u\theta(x_0,u) - \nabla_u\theta(x,u)| &\le \tfrac{1}{3}\,\delta\,g(x,u) \end{aligned}$$

by the continuity of φ, (3.2), and (3.3). Hence

$$\begin{aligned} f(x_0,u,\xi) &= h(x_0,u,\xi) + \varphi(x_0)\cdot\xi^T\nabla_u\theta(x_0,u)\\ &= f(x,u,\xi) + h(x_0,u,\xi) - h(x,u,\xi)\\ &\quad + [\varphi(x_0)-\varphi(x)]\cdot\xi^T\nabla_u\theta(x,u) + \varphi(x_0)\cdot\xi^T[\nabla_u\theta(x_0,u) - \nabla_u\theta(x,u)]\\ &\le f(x,u,\xi) + \tfrac{1}{3}\,\delta(1+g(x,u)|\xi|) + \tfrac{2}{3}\,\delta\,g(x,u)|\xi| \end{aligned}$$

which is (3.2), and where we have used (H_7) and the fact that $|\varphi(x_0)| \le 1$.

Finally, condition (H_3) is replaced by the condition

$$g(x,u)|\xi|(1-|\varphi(x)|) \le f(x,u,\xi) \le 2C\,g(x,u)(1+|\xi|), \tag{3.10}$$

which follows from (3.1) and (H_7). Although (3.10) is weaker than condition (H_3) in [FM2], the proof there carries out even with (3.10). Indeed, condition (H_3) was used in [FM2] only to show that

$$\mathcal{F}(u,\Omega) \geq \int_\Omega f(x,u,\nabla u)\,dx + \int_\Omega f^\infty(x,u,dC(u)) + \int_{S(u)\cap\Omega} K_f(x,u^-,u^+,\nu_u)\,dH_{N-1}(x).$$

The proof of this inequality relies on the blow-up argument introduced in [FM1] which is a *local* argument, in the sense that in order to prove the three main pointwise inequalities (2.10)–(2.11) in [FM2] at points $x_0 \in \Omega$, one is only interested in what happens in a ball $B(x_0,\varepsilon)$. Since in our case $|\varphi(x_0)| < \epsilon_0 < 1$ for some $\epsilon_0 > 0$, if we take ε sufficiently small we can assume that $|\varphi(x)| \leq \epsilon_0$ for all $x \in B(x_0,\varepsilon)$ and thus (3.10) reduces to

$$g(x,u)|\xi|(1-\epsilon_0) \leq f(x,u,\xi) \leq 2C\,g(x,u)(1+|\xi|)$$

for all $(x,u,\xi) \in B(x_0,\varepsilon)\times\mathbb{R}^p\times\mathbb{M}^{p\times N}$, which is the *local* version of (H_3) in [FM2].

In conclusion, we may apply Theorem 2.16 of [FM2] (see Remark 3.10 below) to obtain that for $u \in BV(\Omega;\mathbb{R}^p)$

$$\mathcal{F}(u,\Omega) = \int_\Omega \left\{h(x,u,\nabla u) + \varphi\cdot\nabla u^T\nabla_u\theta(x,u)\right\}\,dx + \int_\Omega h^\infty(x,u,dC(u)) + \int_\Omega (\varphi\otimes\nabla_u\theta(x,u))\cdot\,dC^T(u) + \int_{S(u)\cap\Omega} K_f(x,u^-,u^+,\nu_u)\,dH_{N-1},$$

where

$$K_f(x,a,b,\nu) = \inf\Big\{\int_{Q_\nu}\big[h^\infty(x,\psi(y),\nabla\psi(y)) + \varphi(x)\cdot\nabla\psi^T(y)\nabla_u\theta(x,\psi(y))\big]dy : \psi\in\mathcal{A}(a,b,\nu)\Big\}.$$

Given any $\psi \in \mathcal{A}(a,b,\nu)$ we have

$$\int_{Q_\nu}\varphi(x)\cdot\nabla\psi^T(y)\nabla_u\theta(x,\psi(y))\,dy = \int_{\partial Q_\nu}\varphi(x)\cdot n(y,Q_\nu)\,\theta(x,T\,\psi(y))\,dH_{N-1}(y) = \varphi(x)\cdot(\theta(x,b)-\theta(x,a))\,\nu,$$

and so

$$K_f(x,a,b,\nu) = K_h(x,a,b,\nu) + \varphi(x)\cdot(\theta(x,b)-\theta(x,a))\,\nu.$$

If we now use Lemmas 2.1, 3.7, and 3.8, we finally obtain that $\mathcal{H}(u,\Omega) = \mathcal{L}(u,\Omega)$. This concludes the proof of Theorem 3.2. □

Remark 3.9. The continuity hypotheses (H_1), (H_4), and (3.3) may be replaced by

(H_1)′ h is Carathéodory;

(H_4)′ for all $(x_0,u_0)\in\Omega\times\mathbb{R}^p$ and for all $\delta>0$ there exists $\varepsilon>0$ such that

$$|h(x,u_1,\xi) - h(x,u_2,\xi)| \leq \varepsilon(1+|\xi|)$$

for all $x \in \Omega$ with $|x - x_0| \le \varepsilon$, u_1, $u_2 \in B(u_0, \varepsilon)$, and $\xi \in \mathbb{M}^{p\times N}$, provided $\xi \mapsto f(x, u, \cdot)$ is coercive (e.g., if $g(x, u) \ge \alpha > ||\nabla_u \theta||_{L^\infty}$ for some $\alpha > 0$). In this case, in Theorem 3.2 we would use the integral representation obtained by Bouchitté, Fonseca, and Mascarenhas [BFM] in place of the corresponding result by Fonseca and Müller [FM2].

Proof of Theorem 3.5. By Lemma 3.8 it is enough to find an integral representation for the corresponding $\mathcal{F}_b(u, \Omega)$ in $BV(\Omega; \mathbb{R}^p)$. Let $f_\varepsilon(x, u, \xi) := f(x, u, \xi) + \varepsilon\,|\xi|$, for $\varepsilon \in (0, 1)$, where, as before, $f(x, u, \xi) := h(x, u, \xi) + \varphi(x) \cdot \xi^T \nabla_u \theta(x, u)$, and define

$$\mathcal{F}_\varepsilon(u, \Omega) := \inf_{\{u_n\}} \left\{ \liminf_{n\to\infty} F_\varepsilon(u_n, \Omega) : u_n \in W^{1,1}(\Omega; \mathbb{R}^p),\ u_n \to u \quad \text{in} \quad L^1(\Omega; \mathbb{R}^p) \right\},$$

where

$$F_\varepsilon(u, \Omega) := \int_\Omega f_\varepsilon(x, u(x), \nabla u(x))\, dx, \qquad u \in W^{1,1}(\Omega; \mathbb{R}^p).$$

We claim that

$$\lim_{\varepsilon\to 0} \mathcal{F}_\varepsilon(u, \Omega) = \mathcal{F}_b(u, \Omega).$$

Fix $u \in L^1(\Omega; \mathbb{R}^p)$. For any given $\delta > 0$ there exists a sequence $\{u_n\} \subset W^{1,1}(\Omega; \mathbb{R}^p)$, with $\sup_n ||u_n||_{W^{1,1}} = M < \infty$, such that $u_n \to u$ in $L^1(\Omega; \mathbb{R}^p)$ and

$$\mathcal{F}_b(u, \Omega) + \delta \ge \lim_{n\to\infty} \int_\Omega f(x, u_n(x), \nabla u_n(x))\, dx.$$

In turn, for all $\varepsilon > 0$,

$$\mathcal{F}_b(u, \Omega) + \delta \ge \liminf_{n\to\infty} \int_\Omega f_\varepsilon(x, u_n(x), \nabla u_n(x))\, dx - \varepsilon\, M,$$

and using the definition of $\mathcal{F}_\varepsilon(u, \Omega)$ we obtain

$$\mathcal{F}_b(u, \Omega) + \delta \ge \mathcal{F}_\varepsilon(u, \Omega) - \varepsilon\, M.$$

Therefore

$$\limsup_{\varepsilon\to 0} \mathcal{F}_\varepsilon(u, \Omega) \le \mathcal{F}_b(u, \Omega) + \delta,$$

and it suffices to let $\delta \to 0$ to conclude that

$$\limsup_{\varepsilon\to 0} \mathcal{F}_\varepsilon(u, \Omega) \le \mathcal{F}_b(u, \Omega).$$

Conversely, fix $u \in L^1(\Omega; \mathbb{R}^p)$ and $\varepsilon > 0$. Then there exists a sequence $\{u_n^\varepsilon\} \subset W^{1,1}(\Omega; \mathbb{R}^p)$ such that $u_n^\varepsilon \to u$ in $L^1(\Omega; \mathbb{R}^p)$ as $n \to \infty$ and

(3.11)

$$\mathcal{F}_\varepsilon(u, \Omega) + \varepsilon \ge \lim_{n\to\infty} \int_\Omega \left[f(x, u_n^\varepsilon, \nabla u_n^\varepsilon) + \varepsilon |\nabla u_n^\varepsilon| \right] dx \ge \liminf_{n\to\infty} \int_\Omega f(x, u_n^\varepsilon, \nabla u_n^\varepsilon)\, dx.$$

Without loss of generality we can assume that $\mathcal{F}_\varepsilon(u, \Omega) < \infty$. Since $|\varphi(x)| < 1$ in Ω, by (3.7) we have

$$f(x, u, \xi) = h(x, u, \xi) + \varphi(x) \cdot \xi^T \nabla_u \theta(x, u) \ge 0;$$

hence by (3.11) it follows that $\sup_n ||u_n^\varepsilon||_{W^{1,1}} < \infty$ and so

$$\mathcal{F}_\varepsilon(u,\Omega) + \varepsilon \geq \mathcal{F}_b(u,\Omega).$$

We conclude that

$$\liminf_{\varepsilon\to 0} \mathcal{F}_\varepsilon(u,\Omega) \geq \mathcal{F}_b(u,\Omega)$$

and the claim is proven.

It is not difficult to show that the function $f_\varepsilon(x,u,\xi)$ satisfies conditions (H_1)–(H_5). We omit the details since the proof is very similar to that of Theorem 3.2.

By Theorem 2.16 of [FM2] we obtain that for $u \in BV(\Omega;\mathbb{R}^p)$

$$\begin{aligned}\mathcal{F}_\varepsilon(u,\Omega) = &\int_\Omega \{f(x,u,\nabla u) + \varepsilon|\nabla u|\}\, dx + \int_\Omega f^\infty(x,u,dC(u)) + \varepsilon\int_\Omega |dC(u)| \\ &+ \int_{S(u)\cap\Omega} K_{f_\varepsilon}(x,u^-,u^+,\nu_u)\, dH_{N-1}.\end{aligned} \tag{3.12}$$

If we let $\varepsilon \to 0$ in (3.12) we obtain that

$$\begin{aligned}\lim_{\varepsilon\to 0}\mathcal{F}_\varepsilon(u,\Omega) = &\int_\Omega f(x,u,\nabla u)\, dx + \int_\Omega f^\infty(x,u,dC(u)) \\ &+ \int_{S(u)\cap\Omega} K_f(x,u^-,u^+,\nu_u)\, dH_{N-1},\end{aligned}$$

where we have used the fact that

$$\lim_{\varepsilon\to 0}\int_{S(u)\cap\Omega} K_{f_\varepsilon}(x,u^-,u^+,\nu_u)\, dH_{N-1} = \int_{S(u)\cap\Omega} K_f(x,u^-,u^+,\nu_u)\, dH_{N-1}. \tag{3.13}$$

This convergence follows easily from the Lebesgue dominated convergence theorem and because we may find a constant C_1 independent of ε such that (see the proof of Lemma 2.15 in [FM2])

$$0 \leq K_{f_\varepsilon}(x,a,b,\nu) \leq C_1|a-b|.$$

This concludes the proof of the theorem. □

Remark 3.10. In the proof of Theorem 2.16 of [FM2] the inequality

$$\mathcal{F}(u,S(u)\cap\Omega) \leq \int_{S(u)\cap\Omega} K_f(x,u^-,u^+,\nu_u)\, dH_{N-1}(x) \tag{3.14}$$

was derived by using a result of [FR] which requires the function $f(x,u,\cdot)$ to be coercive, that is, to satisfy the inequality

$$f(x,u,\xi) \geq c_1|\xi| - c_2 \tag{3.15}$$

for all $(x,u,\xi) \in \Omega\times\mathbb{R}^p\times\mathbb{M}^{p\times N}$, which is stronger than condition (H_3). To circumvent this difficulty consider the function $f_\varepsilon(x,u,\xi)$ defined as in the proof of Theorem 3.5. Since it satisfies conditions (H_1)–(H_5) and (3.15), the inequality (3.14) holds for f_ε. Also the inequality $f \leq f_\varepsilon$ clearly implies that

$$\mathcal{F}(u,S(u)\cap\Omega) \leq \mathcal{F}_\varepsilon(u,S(u)\cap\Omega) \leq \int_{S(u)\cap\Omega} K_{f_\varepsilon}(x,u^-,u^+,\nu_u)\, dH_{N-1}(x).$$

If we now let $\varepsilon \to 0$ and use (3.13), we conclude that (3.14) holds also for f.

4. Mesoscopic scale. We are interested in the following constrained minimization problem:

$$\inf\left\{\mathcal{H}(u,\Omega)+\int_\Omega \tau(x,u(x))\,dx:\ u\in BV(\Omega;\mathbb{R}^p),\ u(x)\in K\quad \text{for } \mathcal{L}^N \text{ a.e. } x\in\Omega\right\},$$

where K is a nonempty compact set of $\mathbb{R}^p$, and $\tau:\Omega\times K\to\mathbb{R}$ is a Carathéodory function such that

$$|\tau(x,u)|\le a_0(x)\quad \text{for } \mathcal{L}^N \text{ a.e. } x\in\Omega \text{ and for all } u\in K, \tag{4.1}$$

for some function $a_0\in L^1(\Omega;\mathbb{R})$. In applications in phase transitions, often $K=\{a,b\}$ or K is a convex set.

For $u\in L^1(\Omega;\mathbb{R}^p)$ we define the functional

$$\mathcal{I}(u,\Omega):=\mathcal{H}(u,\Omega)+I_K(u,\Omega),$$

where

$$I_K(u,\Omega):=\begin{cases}\int_\Omega \tau(x,u(x))\,dx & \text{if } u(x)\in K \quad \text{for } \mathcal{L}^N \text{ a.e. } x\in\Omega,\\ +\infty & \text{otherwise.}\end{cases}$$

LEMMA 4.1. *If* (H_1)–(H_7) *hold, then the functional* $\mathcal{I}(u,\Omega)$ *is lower semicontinuous in* $L^1(\Omega;\mathbb{R}^p)$.

Proof. Consider u_n, $u\in L^1(\Omega;\mathbb{R}^p)$ such that $u_n\to u$ in $L^1(\Omega;\mathbb{R}^p)$. If $\liminf_{n\to\infty}\mathcal{I}(u_n,\Omega)=\infty$ there is nothing to prove. Assume that $\liminf_{n\to\infty}\mathcal{I}(u_n,\Omega)<\infty$ and take a subsequence $\{u_{n_k}\}$ which converges pointwise to u for $\mathcal{L}^N$ a.e. $x\in\Omega$, and such that

$$\lim_{k\to\infty}\mathcal{I}(u_{n_k},\Omega)=\liminf_{n\to\infty}\mathcal{I}(u_n,\Omega)<\infty.$$

For k sufficiently large we can assume that $\mathcal{I}(u_{n_k},\Omega)<\infty$; hence

$$\mathcal{I}(u_{n_k},\Omega)=\mathcal{H}(u_{n_k},\Omega)+\int_\Omega \tau(x,u_{n_k}(x))\,dx$$

and $u_{n_k}(x)\in K$ for $\mathcal{L}^N$ a.e. $x\in\Omega$. Since $\{u_{n_k}\}$ converges pointwise to u for $\mathcal{L}^N$ a.e. $x\in\Omega$, we obtain that $u(x)\in K$ for $\mathcal{L}^N$ a.e. $x\in\Omega$. In turn $\mathcal{I}(u,\Omega)=\mathcal{H}(u,\Omega)+\int_\Omega\tau(x,u(x))\,dx$. The assertion now follows from the lower semicontinuity of $\mathcal{H}(u,\Omega)$ in $L^1(\Omega;\mathbb{R}^p)$ and the fact that

$$\lim_{k\to\infty}\int_\Omega \tau(x,u_{n_k}(x))\,dx=\int_\Omega\tau(x,u(x))\,dx$$

by (4.1) and by the Lebesgue dominated convergence theorem. □

In addition to conditions (H_1)–(H_7) we now assume the following hypotheses:

(F_1) There exist a function $\rho\in C(\overline{\Omega}\times\mathbb{R}^p;\mathbb{R}^p)\cap C^1(\Omega\times\mathbb{R}^p;\mathbb{R}^p)$ and a function $b\in L^1(\Omega;\mathbb{R})$ such that

$$|\nabla_x\rho(x,u)|\le b(x)\quad \text{for } \mathcal{L}^N \text{ a.e. } x\in\Omega \text{ and for all } u\in K, \tag{4.2}$$

and

$$h^\infty(x,u,\xi)\ge|\nabla_u\rho(x,u)|\,|\xi| \tag{4.3}$$

for all $(x,u,\xi) \in \Omega \times K \times \mathbb{M}^{p\times N}$;

(F_2) for $\mathcal{L}^N$ a.e. $x \in \Omega$ the function $\rho(x,\cdot) : K \Subset \mathbb{R}^p \to \rho(x,K) \Subset \mathbb{R}^p$ is invertible and $(x,y) \mapsto (\rho(x,\cdot))^{-1}(y)$ is Carathéodory. In addition, there exists a function $c \in L^1(\Omega;\mathbb{R})$ such that

$$|\rho(x,\cdot)^{-1}(v)| \le c(x) \quad \text{for } \mathcal{L}^N \text{ a.e. } x \in \Omega \text{ and for all } v \in \rho(x,K). \tag{4.4}$$

Let

$$D(\mathcal{I}) := \{u \in L^1(\Omega;\mathbb{R}^p) : \mathcal{I}(u,\Omega) < \infty\}.$$

Then

$$D_1 := \left\{u \in BV(\Omega;\mathbb{R}^p) : u(x) \in K \quad \text{for } \mathcal{L}^N \text{ a.e. } x \in \Omega\right\} \subset D(\mathcal{I})$$

but in general the two sets do not coincide, unless one assumes that $h(x,u,\cdot)$ is coercive.

THEOREM 4.2. *There exists a function $u \in D(\mathcal{I})$ such that*

$$\mathcal{I}(u,\Omega) \le \inf\left\{\mathcal{I}(w,\Omega) : w \in D_1\right\}.$$

Proof. Let $\{u_n\} \subset D_1$ be a minimizing sequence; that is,

$$\lim_{n\to\infty} \mathcal{I}(u_n,\Omega) = \inf\left\{\mathcal{I}(w,\Omega) : w \in D_1\right\} < M < \infty.$$

Then, for n sufficiently large,

$$\mathcal{I}(u_n,\Omega) = \mathcal{H}(u_n,\Omega) + \int_\Omega \tau(x,u_n(x))\,dx \le M. \tag{4.5}$$

We claim that $T\,u_n(x) \in K$ for H_{N-1} a.e. $x \in \partial\Omega$. Indeed let $E_n := \{x \in \partial\Omega : T\,u_n(x) \notin K\}$ and suppose for contradiction that $H_{N-1}(E_n) > 0$. Take $x_0 \in E_n$ for which (cf. [Z, Thm. 5.14.4])

$$\lim_{r\to 0} \frac{1}{\mathcal{L}^N(B(x_0,r)\cap\Omega)} \int_{B(x_0,r)\cap\Omega} |u_n(x) - T\,u_n(x_0)|^{N/(N-1)} dx = 0.$$

Since K is compact we have $\operatorname{dist}(T\,u_n(x_0),K) = \varepsilon_0 > 0$, while from the fact that $u_n(x) \in K$ for $\mathcal{L}^N$ a.e. $x \in \Omega$, it follows that

$$\varepsilon_0^{N/(N-1)} \le |u_n(x) - T\,u_n(x_0)|^{N/(N-1)}$$

for $\mathcal{L}^N$ a.e. $x \in B(x_0,r)\cap\Omega$. Taking the average over $B(x_0,r)\cap\Omega$ and letting $r \to 0$, we get a contradiction. Therefore the claim holds, and by (4.5), Theorem 3.2, and (3.5) we have

$$\begin{aligned}\int_\Omega h(x,u_n,\nabla u_n)\,dx + \int_\Omega h^\infty(x,u_n,dC(u_n)) \\ + \int_{S(u_n)\cap\Omega} K_h(x,u_n^-,u_n^+,\nu_{u_n})\,dH_{N-1} \le M_1\end{aligned} \tag{4.6}$$

for some constant M_1 independent of n. By (H_5), (3.1), and (4.6)

$$\begin{aligned}\int_\Omega h^\infty(x,u_n,\nabla u_n)\,dx &\le \int_\Omega (h^\infty(x,u_n,\nabla u_n)-h(x,u_n,\nabla u_n))\,dx + M_1\\ &\le C'||g||_{L^\infty} + C'||g||_{L^\infty}^m \int_\Omega h^{1-m}(x,u_n,\nabla u_n)\,dx + M_1.\end{aligned}$$

Using Hölder's inequality and (4.6) again, we conclude that there exists $M_2 \in (0,\infty)$ such that for all n,

$$\begin{aligned}\int_\Omega h^\infty(x,u_n,\nabla u_n)\,dx + \int_\Omega h^\infty(x,u_n,dC(u_n))\\ + \int_{S(u_n)\cap\Omega} K_h(x,u_n^-,u_n^+,\nu_{u_n})\,dH_{N-1} \le M_2.\end{aligned} \tag{4.7}$$

Define $v_n := \rho(x,u_n(x))$. As in Lemma 3.7, we can show that $v_n(x) \in BV(\Omega;\mathbb{R}^p)$ with

$$Dv_n = \begin{cases} \nabla_x\rho(x,u_n)\,\mathcal{L}^N + \nabla_u\rho(x,u_n)Du_n & \text{on } \Omega\backslash S(u_n),\\ (\rho(x,u_n^+)-\rho(x,u_n^-))\otimes\nu_{u_n}\,H_{N-1}\lfloor S(u_n) & \text{on } S(u_n).\end{cases} \tag{4.8}$$

Furthermore

$$\begin{aligned}\int_\Omega |Dv_n| = \int_\Omega |\nabla v_n|\,dx + \int_\Omega |\nabla_u\rho(x,u_n)\,dC(u_n)|\\ + \int_{S(v_n)\cap\Omega} |(v_n^+ - v_n^-)\otimes\nu_{v_n}|dH_{N-1}.\end{aligned} \tag{4.9}$$

By Remark 2.17 in [FM2] and the fact that $S(v_n) = S(u_n)$ and $\nu_{v_n} = \nu_{u_n}$, we can rewrite the last integral as

$$\int_{S(u_n)\cap\Omega} K_{|\cdot|}(x,v_n^-,v_n^+,\nu_{u_n})\,dH_{N-1},$$

where

$$K_{|\cdot|}(x,v_n^-,v_n^+,\nu_{u_n}) := \inf\left\{\int_{Q_{\nu_{u_n}}} |\nabla\psi(y)|\,dy : \psi\in\mathcal{A}(v_n^-(x),v_n^+(x),\nu_{u_n})\right\}.$$

Given $\eta \in \mathcal{A}(u_n^-(x),u_n^+(x),\nu_{u_n})$, it is clear that the function $\psi(y) := \rho(x,\eta(y))$ belongs to $\mathcal{A}(v_n^-(x),v_n^+(x),\nu_{u_n})$ and that $\nabla\psi(y) = \nabla_u\rho(x,\eta(y))\nabla\eta(y)$. By (4.3) this implies that

$$K_{|\cdot|}(x, v_n^-, v_n^+, \nu_{u_n})$$
$$\leq \inf\left\{\int_{Q_{\nu_{u_n}}} |\nabla_u \rho(x, \eta(y))|\,|\nabla \eta(y)|\,dy : \eta \in \mathcal{A}(u_n^-(x), u_n^+(x), \nu_{u_n})\right\}$$
$$\leq K_h(x, u_n^-, u_n^+, \nu_{u_n}).$$

Therefore, also by (4.2), (4.3), (4.7), (4.8), and (4.9)

$$\begin{aligned}\int_\Omega |Dv_n| &\leq \int_\Omega |\nabla_x \rho(x, u_n)|\,dx + \int_\Omega |\nabla_u \rho(x, u_n)|\,|\nabla u_n|\,dx \\ &+ \int_\Omega |\nabla_u \rho(x, u_n)|\,|dC(u_n)| \\ &+ \int_{S(u_n)\cap\Omega} K_h(x, u_n^-, u_n^+, \nu_{u_n}) dH_{N-1} \leq ||b||_{L^1} + M_2.\end{aligned} \tag{4.10}$$

Finally, since $v_n(x) \in \rho(x, K)$ for $\mathcal{L}^N$ a.e. $x \in \Omega$, $\rho(x, K)$ is a compact set of $\mathbb{R}^p$, and by (4.10) there exists a subsequence, still denoted $\{v_n\}$, which converges strongly in $L^1(\Omega; \mathbb{R}^p)$ and pointwise almost everywhere to a function $v \in BV(\Omega; \mathbb{R}^p)$ (see [Z, Cor. 5.3.4]), with $v(x) \in \rho(x, K)$ for $\mathcal{L}^N$ a.e. $x \in \Omega$. Define $u(x) := (\rho(x, \cdot))^{-1}(v(x))$. By (F_2) the function u is measurable. Since $u_n(x) = (\rho(x, \cdot))^{-1}(v_n(x))$ it follows that $u_n(x) \to u(x)$ for $\mathcal{L}^N$ a.e. $x \in \Omega$; thus, $u(x) \in K$ for $\mathcal{L}^N$ a.e. $x \in \Omega$. Moreover, by (4.4) we have that $|u_n(x)| \leq c(x)$ for $\mathcal{L}^N$ a.e. $x \in \Omega$; therefore, by the Lebesgue dominated convergence theorem $u_n \to u$ strongly in $L^1(\Omega; \mathbb{R}^p)$. By Lemma 4.1 we conclude that

$$\mathcal{I}(u, \Omega) \leq \inf\{\mathcal{I}(u, \Omega) : u \in D_1\}. \qquad \square$$

COROLLARY 4.3. *Assume that conditions* (F_1) *and* (F_2) *in Theorem* 4.2 *are replaced by the assumption*

$$h^\infty(x, u, \xi) \geq \alpha|\xi| \qquad \text{for all } (x, u, \xi) \in \Omega \times K \times \mathbb{M}^{p\times N}$$

for some $\alpha > 0$. *Then* $D_1 = D(\mathcal{I})$ *and there exists a function* $u \in D_1$ *such that*

$$\mathcal{I}(u, \Omega) = \inf\{\mathcal{I}(w, \Omega) : w \in D_1\}.$$

Proof. It suffices to take $\rho(x, u) := \alpha\, u$ in Theorem 4.2. Then $v_n = \alpha\, u_n$ converge strongly in $L^1(\Omega; \mathbb{R}^p)$ to a function $v \in BV(\Omega; \mathbb{R}^p)$, and therefore $u := \frac{1}{\alpha}\, v$ is the desired minimizer. $\square$

5. Nucleation: The scalar case. In this section we study the constrained minimization problem introduced in section 4, restricted to the scalar case $p = 1$, when K is a closed, connected subset of $\mathbb{R}$ (not necessarily bounded), and when the potential $\tau(x, u)$ is given by

$$\tau(x, u) := \tau_1(x, u) + \psi(x)\,\tau_2(u),$$

where $\tau_1(x, u)$ is a Carathéodory function, concave in the u variable, ψ is a nonnegative, measurable function, and τ_2 is a continuous function such that

(5.1)
all the connected components of $S := \{u \in \text{int } K : \tau_2^{**}(u) < \tau_2(u)\}$ are bounded,

where τ_2^{**} is the convex envelope of τ_2. As remarked in [V2], (5.1) holds if

$$\limsup_{|u|\to\infty} \frac{\tau_2(u)}{|u|} = \infty.$$

Furthermore we assume that

$$\tau(x,u) \geq -L_1 - L_2|u| \quad \text{for } \mathcal{L}^N \text{ a.e. } x \in \Omega \text{ and for all } u \in K \tag{5.2}$$

for some $L_1,\ L_2 > 0$.

Under appropriate assumptions on the functions h and θ, we prove that minimizers $u \in L^1(\Omega;\mathbb{R}^p)$ of

$$\mathcal{I} : v \in L^1(\Omega;K) \mapsto \mathcal{H}(v,\Omega) + \int_\Omega \tau(x,v(x))\,dx$$

have the phase structure

$$u(\Omega) \subset K\backslash S.$$

In particular, if $K=[a,b]$, if τ_2 is concave in $[a,b]$, and if u is a minimizer of $\mathcal{I}$, then u must have a two-phase structure, i.e., there exists a set $\Omega_0 \subset \Omega$ such that $u(x)=a$ for $\mathcal{L}^N$ a.e. $x \in \Omega_0$ and $u(x)=b$ for $\mathcal{L}^N$ a.e. $x \in \Omega\backslash\Omega_0$. This result has important applications in *nucleation phenomena* which have been studied extensively by Visintin in [V1, V2], where usually K is bounded, $\tau_1(x,u) := -\xi(x)\,u$, $\xi \in L^\infty(\Omega;\mathbb{R})$ is proportional to the relative temperature, and $\psi(x)\,\tau_2(u)$ is the double well potential $\psi(x)(b-u)(u-a)$ (see Remark 5.2 below). Given a simple function $u \in L^1(\Omega;K)$ of the form

$$u(x) = \sum_{i=1}^k c_i\,\chi_{\omega_i}(x), \tag{5.3}$$

with $c_i \in K$, $\mathcal{L}^N(\omega_i) > 0$ for all $i=1,\dots,k$, and $\mathcal{L}^N(\Omega\backslash \cup_{i=1}^k \omega_i) = 0$, without loss of generality we may assume that

$$\inf K \leq c_1 < c_2 < \cdots < c_k \leq \sup K. \tag{5.4}$$

THEOREM 5.1. *Let $\mathcal{E}$ be an algebra of measurable subsets of Ω, and consider a functional $\mathcal{V} : L^1(\Omega;\mathbb{R}) \to [0,\infty]$ such that*

$$\mathcal{S}_1 := \left\{ u \in L^1(\Omega;\mathbb{R}) :\ u = \sum_{i=1}^k c_i\,\chi_{\omega_i},\ \omega_i \in \mathcal{E}, k \in \mathbb{N} \right\} \subset D(\mathcal{V}),$$

where

$$D(\mathcal{V}) := \{u \in L^1(\Omega;\mathbb{R}) :\ \mathcal{V}(u) < \infty\},$$

and

(i) *for any $u \in D(\mathcal{V}) \cap L^1(\Omega;K)$ there exists a sequence $\{u_n\} \subset \mathcal{S}_1 \cap L^1(\Omega;K)$ converging to u in $L^1(\Omega;K)$ and such that*

$$\limsup_{n\to\infty} \mathcal{V}(u_n) \leq \mathcal{V}(u).$$

(ii) *For any* $u \in \mathcal{S}_1$ *of the form* (5.3)–(5.4)*, with* $k \geq 2$*, there holds*

$$\mathcal{V}(u) = \sum_{i=1}^{k-1} (c_{i+1} - c_i) \mathcal{V}(\chi_{\cup_{r=i+1}^{k} \omega_r}). \tag{5.5}$$

(iii) *The function* $c \mapsto \mathcal{V}(c)$ *is concave in* K*.*

In addition, suppose that the functional $u \mapsto \int_\Omega \tau(x, u(x))\, dx$ *is continuous in* $D(\mathcal{V}) \cap L^1(\Omega; K)$*. Then*

$$\inf \{\mathcal{V}(u) + I_K(u, \Omega) : u \in D(\mathcal{V})\} = \inf \big\{\mathcal{V}(u) + I_K(u, \Omega) : u \in D(\mathcal{V}),\, u(x) \in K \backslash S \text{ for } \mathcal{L}^N \text{ a.e. } x \in \Omega\big\}.$$

Remark 5.2. (i) The functional $\mathcal{V}(u) + I_K(u, \Omega)$ is well defined by (5.2).

(ii) Theorem 5.1 is closely related to Theorem 2 in [V2], where $K = \mathbb{R}$ and conditions (i) and (ii) are replaced by the assumption that $\mathcal{V}$ satisfies the *generalized co-area formula*

$$\mathcal{V}(u) = \int_{\mathbb{R}} \mathcal{V}(\chi_{\{x \in \Omega:\, u(x) \geq t\}})\, dt. \tag{5.6}$$

It is easy to see that (5.6) reduces to (5.5) for functions u of the form (5.3)–(5.4). Therefore, (5.5) is weaker than (5.6). On the other hand, conditions (i) and (5.6) do not seem to be related. Indeed, consider the functional

$$\mathcal{V}(u) := \int_\Omega |Du| + \begin{cases} \int_\Omega \max\{u(x), 0\} dx & \text{if } H_{N-1}(S(u) \cap \Omega) = 0, \\ 0 & \text{otherwise.} \end{cases}$$

From the proof of Theorem 5.4 below it follows that $\mathcal{V}$ satisfies hypotheses (i)–(iii) of Theorem 5.1. Take $u(x) := 1$ in (5.6); then $\mathcal{V}(1) = \mathcal{L}^N(\Omega)$, while the right-hand side of (5.6) is infinite. Therefore, (5.6) fails. We note that $\mathcal{V}$ is not lower semicontinuous in L^1.

We remark that Theorem 5.1 may be applied to a large class of functionals of the form (1.1), for which the co-area formula might not hold.

(iii) If, in addition to hypotheses (i)–(iii) in Theorem 5.1, we assume that $\mathcal{V}$ is lower semicontinuous in $L^1(\Omega; \mathbb{R})$, that $K = \mathbb{R}$, and that there exists a set $\omega \in \mathcal{E}$ with $0 < \mathcal{L}^N(\omega) < \mathcal{L}^N(\Omega)$, then $\mathcal{V}$ satisfies the following properties:

(1) $\mathcal{V}(c) = 0$ for all $c \in \mathbb{R}$;
(2) $\mathcal{V}(\lambda\, u) = \lambda\, \mathcal{V}(u)$ for all $\lambda > 0$ and $u \in D(\mathcal{V})$;
(3) $\mathcal{V}(u + c) = \mathcal{V}(u)$ for all $c \in \mathbb{R}$ and $u \in D(\mathcal{V})$;
(4) $\mathcal{V}(u) \geq \int_{\mathbb{R}} \mathcal{V}(\chi_{\{x \in \Omega:\, u(x) \geq t\}})\, dt$ for all $u \in D(\mathcal{V})$.

In order to prove the first property, define

$$u_n(x) := \begin{cases} c + \varepsilon_n & \text{if } x \in \omega, \\ c & \text{if } x \in \Omega \backslash \omega, \end{cases}$$

where $\varepsilon_n := \frac{1}{n} \min\{1, 1/\mathcal{V}(\chi_\omega)\}$ if $\mathcal{V}(\chi_\omega) > 0$, and $\varepsilon_n := \frac{1}{n}$ otherwise. Clearly $u_n \to c$ in $L^1(\Omega; \mathbb{R})$; therefore, by the lower semicontinuity of $\mathcal{V}$ and (5.5)

$$0 \leq \mathcal{V}(c) \leq \liminf_{n \to \infty} \mathcal{V}(u_n) = \lim_{n \to \infty} \varepsilon_n \mathcal{V}(\chi_\omega) = 0,$$

where we have used the fact that $\mathcal{V}(\chi_\omega) < \infty$ because $S_1 \subset D(\mathcal{V})$.

We omit the proofs of properties (2) and (3) since they follow quite easily from hypotheses (i) and (ii) of Theorem 5.1 and from the lower semicontinuity of $\mathcal{V}$.

In order to show (4), fix $u \in D(\mathcal{V})$. By (i) there exists a sequence $\{u_n\} \subset \mathcal{S}_1$ converging to u in $L^1(\Omega;\mathbb{R})$ and $\mathcal{L}^N$ a.e. $x \in \Omega$ such that

$$\mathcal{V}(u) \geq \lim_{n\to\infty} \mathcal{V}(u_n) = \lim_{n\to\infty} \int_{\mathbb{R}} \mathcal{V}(\chi_{\{x\in\Omega:\, u_n(x)\geq t\}})\, dt \geq \int_{\mathbb{R}} \liminf_{n\to\infty} \mathcal{V}(\chi_{\{x\in\Omega:\, u_n(x)\geq t\}})\, dt$$

by (5.5) and Fatou's lemma. Since $\mathcal{L}^N(\{x \in \Omega :\, u(x) = t\}) = 0$ for all $t \in \mathbb{R}\backslash M$, where $\mathcal{L}^1(M) = 0$, we fix $t \in \mathbb{R}\backslash M$ and take a subsequence $\{u_{n_k}\}$ of $\{u_n\}$ such that

$$\liminf_{n\to\infty} \mathcal{V}(\chi_{\{x\in\Omega:\, u_n(x)\geq t\}}) = \lim_{k\to\infty} \mathcal{V}(\chi_{\{x\in\Omega:\, u_{n_k}(x)\geq t\}}).$$

Then $\{\chi_{\{x\in\Omega:\, u_{n_k}(x)\geq t\}}\}$ converges pointwise to $\chi_{\{x\in\Omega:\, u(x)\geq t\}}$ for $\mathcal{L}^N$ a.e. $x \in \Omega$ and, by the Lebesgue dominated convergence theorem, also strongly in $L^1(\Omega;\mathbb{R})$. Therefore, by the lower semicontinuity of $\mathcal{V}$,

$$\liminf_{n\to\infty} \mathcal{V}(\chi_{\{x\in\Omega:\, u_n(x)\geq t\}}) \geq \mathcal{V}(\chi_{\{x\in\Omega:\, u(x)\geq t\}})$$

for $\mathcal{L}^1$ a.e. $t \in \mathbb{R}$, and we conclude that

$$\int_{\mathbb{R}} \liminf_{n\to\infty} \mathcal{V}(\chi_{\{x\in\Omega:\, u_n(x)\geq t\}})\, dt \geq \int_{\mathbb{R}} \mathcal{V}(\chi_{\{x\in\Omega:\, u(x)\geq t\}})\, dt.$$

We do not know if the reversed inequality of (4) holds, i.e., if the co-area formula (5.6) is satisfied.

Let

$$\beta := \inf\left\{\mathcal{V}(u) + I_K(u,\Omega) :\, u \in D(\mathcal{V}),\, u(x) \in K\backslash S \quad \text{for } \mathcal{L}^N \text{ a.e. } x \in \Omega\right\}.$$

LEMMA 5.3. *If $u \in \mathcal{S}_1$, then*

$$\mathcal{V}(u) + I_K(u,\Omega) \geq \beta.$$

Proof. As $I_K(u,\Omega) = \infty$ for $u \notin L^1(\Omega;K)$ it suffices to prove the result for $u \in \mathcal{S}_1 \cap L^1(\Omega;K)$. By (5.1) we can decompose the open set S as a disjoint union of bounded intervals

$$S = \cup_{r\in\mathcal{R}}(a_r, b_r).$$

Following Visintin [V2] we replace the function τ_2 by

$$\tilde{\tau}_2(u) := \begin{cases} \tau_2(u) & \text{if } u \in \mathbb{R}\backslash S, \\ \dfrac{\tau_2(b_r) - \tau_2(a_r)}{b_r - a_r}\,(u - a_r) + \tau_2(a_r) & \text{if } u \in (a_r, b_r), \end{cases}$$

and denote by $\tilde{\tau}$ and $\tilde{I}_K(u,\Omega)$ the corresponding functionals. Define

$$\beta_i := \mathcal{V}(\chi_{\cup_{r=i+1}^{k}\omega_r}), \qquad \alpha_i(c) := \int_{\omega_i} \tilde{\tau}(x,c)\, dx.$$

Then by (5.3), (5.4), and (5.5)

$$\mathcal{V}(u) + \tilde{I}_K(u,\Omega) = \sum_{i=1}^{k-1}(c_{i+1} - c_i)\beta_i + \sum_{i=1}^{k} \alpha_i(c_i).$$

Let $r \in \mathcal{R}$ be such that $c_i \in (a_r, b_r)$ for some $i \in \{1, \dots, k\}$. There can only be finitely many such r. Assume that $k \geq 2$, and suppose that $c_l \in (a_r, b_r)$, $l \in \{2, \dots, k-1\}$, $c_i \leq a_r$ for all $i < l$ (the cases where $l = 1$ or $l = k$ can be treated analogously). Define the function

$$\Phi(t) := \sum_{i=1,\, i\neq l-1,\, l}^{k-1} \beta_{i+1}(c_{i+1} - c_i) + \beta_l(t - c_{l-1}) + \beta_{l+1}(c_{l+1} - t) + \sum_{i=1,\, i\neq l}^{k} \alpha_i(c_i) + \alpha_l(t)$$

for $t \in [a_r, d]$, where $d := c_{l+1}$ if $c_{l+1} \leq b_r$ and $d := b_r$ if $c_{l+1} > b_r$. Since $\tau_2(u) \geq \tilde{\tau}_2(u)$ by construction, then clearly $\mathcal{V}(u) + I_K(u, \Omega) \geq \mathcal{V}(u) + \tilde{I}_K(u, \Omega) = \Phi(c_l)$. Observe that since $\tilde{\tau}(x, \cdot) = \tau_1(x, \cdot) + \psi(x)\, \tilde{\tau}_2(\cdot)$ is concave in $[a_r, d]$, then the function $\alpha_l(\cdot)$ is also concave in $[a_r, d]$, and $\Phi(t)$, being the sum of a linear function and a concave function, attains its minimum at one of the endpoints Q of $[a_r, d]$. It follows that

$$\mathcal{V}(u) + I_K(u, \Omega) \geq \mathcal{V}(u) + \tilde{I}_K(u, \Omega) = \Phi(c_l) \geq \Phi(Q) = \mathcal{V}(\bar{u}) + \tilde{I}_K(\bar{u}, \Omega),$$

where

$$\bar{u}(x) := \begin{cases} \sum_{i=1,\, i\neq l}^{k} c_i\, \chi_{\omega_i}(x) + a_r \chi_{\omega_l}(x) & \text{if } Q = a_r, \\ \sum_{i=1,\, i\neq l}^{k} c_i\, \chi_{\omega_i}(x) + b_r \chi_{\omega_l}(x) & \text{if } Q = b_r, \\ \sum_{i=1,\, i\neq l, l+1}^{k} c_i\, \chi_{\omega_i}(x) + c_{l+1} \chi_{\omega_l \cup \omega_{l+1}}(x) & \text{if } Q = c_{l+1}. \end{cases}$$

If $k = 1$, namely, if $u(x) \equiv c$, then

$$\mathcal{V}(u) + I_K(u, \Omega) \geq \mathcal{V}(u) + \tilde{I}_K(u, \Omega) = \mathcal{V}(c) + \int_\Omega \tilde{\tau}(x, c)\, dx. \tag{5.7}$$

Assume that $c \in (a_r, b_r)$ for some $r \in \mathcal{R}$. Since by (iii) in the statement of Theorem 5.1 and by the construction of $\tilde{\tau}_2$ the right-hand side of (5.7) is a concave function of $c \in [a_r, b_r]$, its infimum is attained at one of the endpoints, say, at b_r, and thus we can replace $u(x)$ by $\bar{u}(x) := b_r \in K \setminus S$.

We conclude that it is energetically possible to reduce at least by one the number of values c_i between a_r and b_r. Repeating this procedure for the finite number of intervals (a_r, b_r), which contain at least one of the c_i, by means of a finite induction argument we can construct a simple function $\hat{u}$ of the form

$$\hat{u}(x) = \sum_{i=1}^{\hat{k}} \hat{c}_i\, \chi_{\hat{\omega}_i}(x),$$

where $\hat{k} \leq k$, such that $\hat{u}(\Omega) \subset K \setminus (a_r, b_r)$ for any $r \in \mathcal{R}$ and $\mathcal{V}(u) + I_K(u, \Omega) \geq \mathcal{V}(\hat{u}) + \tilde{I}_K(\hat{u}, \Omega)$. Since $\tau_2(u) = \tilde{\tau}_2(u)$ for $u \in K \setminus S$, it follows that $I_K(\hat{u}, \Omega) = \tilde{I}_K(\hat{u}, \Omega)$ and thus $\mathcal{V}(u) + I_K(u, \Omega) \geq \mathcal{V}(\hat{u}) + I_K(\hat{u}, \Omega) \geq \beta$. This concludes the proof of the lemma. □

Proof of Theorem 5.1. Let $u \in D(\mathcal{V}) \cap L^1(\Omega; K)$. By (i) there exists a sequence $\{u_n\} \subset \mathcal{S}_1 \cap L^1(\Omega; K)$ converging to u in $L^1(\Omega; K)$ such that

$$\limsup_{n\to\infty} \mathcal{V}(u_n) \leq \mathcal{V}(u).$$

Moreover, by hypothesis,

$$\lim_{n\to\infty} \int_\Omega \tau(x, u_n(x))\, dx = \int_\Omega \tau(x, u(x))\, dx,$$

and since by Lemma 5.3 $\mathcal{V}(u_n) + I_K(u_n, \Omega) \geq \beta$, it follows that

$$\mathcal{V}(u) + I_K(u, \Omega) \geq \beta,$$

and we conclude that

$$\inf \{\mathcal{V}(u) + I_K(u, \Omega) : u \in D(\mathcal{V})\} \geq \beta.$$

The reversed inequality is trivially satisfied. □

In order to apply Theorem 5.1 to functionals of the form (1.1), we consider the special case where

$$h = h(x, \xi) \quad \text{is positively homogeneous of degree one in } \xi, \quad \theta(x, u) := \hat{\sigma}\, u, \quad \hat{\sigma} \neq 0,$$

and hypotheses (A_1)–(A_7) in section 2 are satisfied. Clearly $h(x, \xi) = h^\infty(x, \xi)$, and by Theorem 2.2 (see also Corollary 3.3, (3.9), and Lemma 3.8), for $u \in BV(\Omega; \mathbb{R})$ we have

$$\begin{aligned}\mathcal{H}(u, \Omega) = &\int_\Omega f(x, \nabla u)\, dx + \int_\Omega f(x, dC(u)) \\ &+ \int_{S(u)\cap\Omega} f(x, (u^+ - u^-)\nu_u)\, dH_{N-1} + \hat{\sigma} \int_\Omega u \operatorname{div} \varphi\, dx,\end{aligned}$$

where, we recall,

$$f(x, \xi) = h(x, \xi) + \hat{\sigma}\, \varphi(x) \cdot \xi \qquad \text{for all } (x, \xi) \in \Omega \times \mathbb{R}^N.$$

Furthermore, we assume that the potential τ also satisfies the growth condition

$$\tau(x, u) \leq b_1(x) + M_1(1 + |u|^{q_c}) \qquad \text{for } \mathcal{L}^N \text{ a.e. } x \in \Omega \text{ and all } u \in K, \tag{5.8}$$

where $b_1 \in L^1(\Omega, \mathbb{R})$, $M_1 > 0$, and, as before, q_c is the Sobolev exponent $q_c := N/(N-1)$ if $N > 1$ and $q_c < \infty$ if $N = 1$. Define

$$\mathcal{V}(u) := \begin{cases} \int_\Omega f(x, Du) & \text{if } u \in BV(\Omega; \mathbb{R}), \\ \infty & \text{otherwise,} \end{cases}$$

where

$$\int_\Omega f(x, Du) := \int_\Omega f(x, \nabla u)\, dx + \int_\Omega f(x, dC(u)) + \int_{S(u)\cap\Omega} f(x, (u^+ - u^-)\nu_u)\, dH_{N-1},$$

and take $\mathcal{E}$ to be the algebra generated by the class of open polyhedral subsets of Ω. If $u \in S_1 \cap L^1(\Omega; K)$ has the form (5.3), then either $\omega_i = E_i$ or $\omega_i = \Omega \backslash E_i$, where E_i is an open polyhedral set of Ω. Therefore, in both cases $\partial\omega_i \cap \Omega = \partial E_i \cap \Omega$, which is given by the intersection of a finite union of hyperplanes. Consequently, if $i < j$ for H_{N-1} a.e. $x \in \partial\omega_j \cap \partial\omega_i$, the outward unit normal $\nu_j(x)$ to the set ω_j at the point x coincides with $-\nu_i(x)$. Moreover, for $j < i < l$ we have

$$H_{N-1}(\partial\omega_j \cap \partial\omega_i \cap \partial\omega_l \cap \Omega) = 0 \tag{5.9}$$

and

(5.10)

$$\begin{aligned}&\Omega = \cup_{i=1}^k \overline{\omega}_i \cap \Omega, \qquad \partial\omega_i \cap \Omega = \cup_{j\neq i} \partial\omega_i \cap \partial\omega_j \cap \Omega, \\ &\partial\left(\cup_{l=i}^k \overline{\omega}_l \cap \Omega\right) = \partial\left(\cup_{l=i+1}^k \overline{\omega}_l \cap \Omega\right) \setminus \left(\cup_{l=i+1}^k \partial\omega_l \cap \partial\omega_i \cap \Omega\right) \cup \left(\cup_{j=1}^{i-1} \partial\omega_j \cap \partial\omega_i \cap \Omega\right).\end{aligned}$$

These properties will be useful in the sequel.

The main result of the section is the following theorem.

THEOREM 5.4. *If* (H_1)–(H_7) *and* (5.8) *are verified, then*

$$\inf\{\mathcal{I}(u,\Omega):\ u\in BV(\Omega;K)\}=\inf\{\mathcal{I}(u,\Omega):\ u\in BV(\Omega;K\backslash S)\}.$$

Remark 5.5. Theorem 5.4 no longer holds in general if $\theta(x,\cdot)$ is nonlinear. Indeed, consider the simple case where $\Omega:=(c,d)$, $K:=[-1,1]$,

$$\mathcal{H}(u,\Omega):=\sigma\int_\Omega|Du(x)|-\int_{\partial\Omega}\sin(\pi\,T\,u(x))\,dH_{N-1}(x),\quad u\in BV(\Omega;\mathbb{R}),$$

and $\tau(x,u):=a(1-u^2)$. Here $K\setminus S=\{-1,1\}$, and if $\sigma>\pi$, then all conditions (H_1)–(H_7) are satisfied. Let $u\in BV(\Omega;\mathbb{R})$, with $u(x)\in\{-1,1\}$ for $\mathcal{L}^N$ a.e. $x\in\Omega$. Then $\mathcal{I}(u,\Omega)=\sigma\int_\Omega|Du(x)|\geq 0$. On the other hand, if we take $\bar u(x)\equiv\frac12$, then $\mathcal{I}(\bar u,\Omega)=-2+\frac34\,a(d-c)<0$ provided $a(d-c)<\frac83$.

For the proof of the lemma below we refer to [F], [LM], and [Re].

LEMMA 5.6. *Let* $f:\Omega\times\mathbb{R}^N\to\mathbb{R}$ *be a continuous function such that* $\xi\in\mathbb{R}^N\mapsto f(x,\xi)$ *is positively homogeneous of degree one for all* $x\in\Omega$, *and*

$$0\leq f(x,\xi)\leq C(1+|\xi|)\quad \textit{for some } C>0,\ \textit{all } x\in\Omega \textit{ and } \xi\in\mathbb{R}^N.$$

Let $\{\mu_n\}$ *be a sequence of Radon measures converging weakly-*$\star$ *to a Radon measure* μ *and*

$$\lim_{n\to\infty}|\mu_n|(\Omega)=|\mu|(\Omega).$$

Then

$$\lim_{n\to\infty}\int_\Omega f(x,d\mu_n)=\int_\Omega f(x,d\mu).$$

Proof of Theorem 5.4. We claim that $\mathcal{V}$ satisfies conditions (i)–(iii) of Theorem 5.1. To prove (i) fix $u\in BV(\Omega;K)$. We can find a sequence $\{u_n\}$ of the form

$$u_n(x)=\sum_{i=1}^{k_n}c_{n,i}\,\chi_{\omega_{n,i}}(x)$$

such that u_n converges strongly to u in $L^1(\Omega;\mathbb{R})$ and $\int_\Omega|Du_n|\to\int_\Omega|Du|$ (see [AMT]). Here $c_{n,i}\in\mathbb{R}$, the sets $\omega_{n,i}$ are open polyhedral set of Ω, $\cup_{i=1}^{k_n}\overline{\omega}_{n,i}\cap\Omega=\Omega$, and as in (5.4)

$$c_{n,1}<c_{n,2}<\cdots<c_{n,k_n}.$$

By (5.10) it is not difficult to see that

$$S(u_n)=\cup_{(i,j)\in J_n}(\partial\omega_{n,i}\cap\partial\omega_{n,j}),$$

where $J_n=\{(i,j)\in\mathbb{N}^2:\ 1\leq i<j\leq k_n\}$. Furthermore, if $x_0\in\partial\omega_{n,i}\cap\partial\omega_{n,j}$, then $u_n^+(x_0)=c_{n,j}$, $u_n^-(x_0)=c_{n,i}$, and

$$\int_\Omega|Du_n|=\sum_{(i,j)\in J_n}(c_{n,j}-c_{n,i})H_{N-1}(\partial\omega_i\cap\partial\omega_j\cap\Omega).$$

Let $\bar{u}_n(x) := \sum_{i=1}^{k_n} d_{n,i}\,\chi_{\omega_{n,i}}(x)$, where

$$d_{n,i} = \begin{cases} \sup K & \text{if } c_{n,i} \geq \sup K, \\ c_{n,i} & \text{if } -\inf K < c_{n,i} < \sup K, \\ \inf K & \text{if } c_{n,i} \leq \inf K. \end{cases}$$

Since $0 \leq (d_{n,j} - d_{n,i}) \leq (c_{n,j} - c_{n,i})$, it follows that $\int_\Omega |D\bar{u}_n| \leq \int_\Omega |Du_n|$. Consequently

$$\limsup_{n\to\infty} \int_\Omega |D\bar{u}_n| \leq \int_\Omega |Du|.$$

On the other hand, since $u(x) \in K$ for $\mathcal{L}^N$ a.e. $x \in \Omega$, then $|u(x) - \bar{u}_n(x)| \leq |u(x) - u_n(x)|$ by construction, and so $\{\bar{u}_n\}$ converges strongly to u in $L^1(\Omega;\mathbb{R})$. By the lower semicontinuity of the total variation we have that $\int_\Omega |Du| \leq \liminf_{n\to\infty} \int_\Omega |D\bar{u}_n|$; thus

$$\lim_{n\to\infty} \int_\Omega |D\bar{u}_n| = \int_\Omega |Du|$$

and from Lemma 5.6 we conclude that

$$\lim_{n\to\infty} \int_\Omega f(x, D\bar{u}_n) = \int_\Omega f(x, Du).$$

In order to verify (ii) in the statement of Theorem 5.1, fix $u \in BV(\Omega;\mathbb{R})$ of the form (5.3)–(5.4), where $c_i \in K$, $\omega_i \in \mathcal{E}$, $\cup_{i=1}^k \overline{\omega}_i \cap \Omega = \Omega$, $k \geq 2$, and

$$c_1 < c_2 < \cdots < c_k.$$

Since by homogeneity $f(x,0) = 0$, we have

$$\mathcal{V}(u) = \sum_{(i,j)\in J} (c_j - c_i) \int_{\partial\omega_i \cap \partial\omega_j \cap \Omega} f(x, \nu_j)\, dH_{N-1}$$

with $J = \{(i,j) \in \mathbb{N}^2 : 1 \leq i < j \leq k\}$, or, equivalently,

(5.11)

$$\begin{aligned} \mathcal{V}(u) = {} & c_k \sum_{j=1}^{k-1} \int_{\partial\omega_k \cap \partial\omega_j \cap \Omega} f(x, \nu_k)\, dH_{N-1} - c_1 \sum_{l=2}^{k} \int_{\partial\omega_l \cap \partial\omega_1 \cap \Omega} f(x, \nu_l)\, dH_{N-1} \\ & + \sum_{i=2}^{k-1} c_i \left(\sum_{j=1}^{i-1} \int_{\partial\omega_i \cap \partial\omega_j \cap \Omega} f(x, \nu_i)\, dH_{N-1} - \sum_{l=i+1}^{k} \int_{\partial\omega_l \cap \partial\omega_i \cap \Omega} f(x, \nu_l)\, dH_{N-1} \right). \end{aligned}$$

By (5.9) and (5.10) we can rewrite the first two terms as, respectively,

$$c_k \int_{\partial\omega_k \cap \Omega} f(x, \nu_k)\, dH_{N-1} \quad \text{and} \quad -c_1 \int_{\partial(\cup_{l=2}^k \overline{\omega_l}) \cap \Omega} f(x, -\nu_1)\, dH_{N-1},$$

and for $i \in \{2, \ldots, k-1\}$

(5.12)

$$\begin{aligned} & \sum_{j=1}^{i-1} \int_{\partial\omega_i \cap \partial\omega_j \cap \Omega} f(x, \nu_i)\, dH_{N-1} - \sum_{l=i+1}^{k} \int_{\partial\omega_l \cap \partial\omega_i \cap \Omega} f(x, \nu_l)\, dH_{N-1} \\ & \qquad = \int_{\partial(\cup_{l=i}^k \overline{\omega_l}) \cap \Omega} f(x, \hat{\nu}_i)\, dH_{N-1} - \int_{\partial(\cup_{l=i+1}^k \overline{\omega_l}) \cap \Omega} f(x, \hat{\nu}_{i+1})\, dH_{N-1}, \end{aligned}$$

where $\hat{\nu}_i$ and $\hat{\nu}_{i+1}$ are, respectively, the outward unit normals to the sets $\cup_{l=i}^{k} \overline{\omega_l}$ and $\cup_{l=i+1}^{k} \overline{\omega_l}$. It now follows from (5.11)–(5.12) that

$$\mathcal{V}(u) = c_k \mathcal{V}(\chi_{\omega_k}) - c_1 \mathcal{V}(\chi_{\cup_{l=2}^{k}\omega_l}) + \sum_{i=2}^{k-1} c_i \mathcal{V}(\chi_{\cup_{l=i}^{k}\omega_l}) - \sum_{i=2}^{k-1} c_i \mathcal{V}(\chi_{\cup_{l=i+1}^{k}\omega_l}),$$

which is (ii) in the statement of Theorem 5.1.

Finally, by (5.2), (5.8), and the Sobolev inequality, the functional $u \mapsto \int_\Omega \tau(x, u(x))\, dx$ is continuous in $BV(\Omega; K)$ (see [K, Thm. 2.1]). Therefore, we can now apply Theorem 5.1 (with $\tau(x, u)$ replaced by $\tau(x, u) + \hat{\sigma}\, u \operatorname{div} \varphi(x)$) to obtain that

$$\inf \{\mathcal{I}(u, \Omega) : u \in BV(\Omega; K)\} = \inf \{\mathcal{I}(u, \Omega) : u \in BV(\Omega; K \backslash S)\}. \qquad \square$$

COROLLARY 5.7. *Assume that $K = [a, b]$ in Theorem 5.4. Then there exists a function $u \in BV(\Omega; [a, b]\backslash S)$ such that*

$$\mathcal{I}(u, \Omega) = \inf \{\mathcal{I}(u, \Omega) : u \in BV(\Omega; [a, b])\}.$$

Proof. By Theorem 5.4,

$$\inf \{\mathcal{I}(u, \Omega) : u \in BV(\Omega; [a, b])\} = \inf \{\mathcal{I}(u, \Omega) : u \in BV(\Omega; [a, b]\backslash S)\} = \beta.$$

To complete the proof it suffices to apply Corollary 4.3, with $K := [a, b]\backslash S$, to find $u \in BV(\Omega; [a, b]\backslash S)$ such that $\mathcal{I}(u, \Omega) = \beta$. $\square$

Remark 5.8. If we assume that τ_2 is concave in $[a, b]$, then $S = (a, b)$ and consequently the minimizer u in Corollary 5.7 has the property that $u(x) \in \{a, b\}$ for $\mathcal{L}^N$ a.e. $x \in \Omega$.

Acknowledgment. The authors are deeply indebted to Luc Tartar for many fruitful and stimulating discussions on the subject of this work.

REFERENCES

[ABS] G. ALBERTI, G. BOUCHITTÉ, AND P. SEPPECHER, *Phase transition with line tension effect*, Arch. Rational Mech. Anal., to appear.

[ADM1] L. AMBROSIO AND G. DAL MASO, *A general chain rule for distributional derivatives*, Proc. Amer. Math. Soc., 108 (1988), pp. 691–702.

[ADM2] L. AMBROSIO AND G. DAL MASO, *On the relaxation in $BV(\Omega; \mathbb{R}^m)$ of quasi-convex integrals*, J. Funct. Anal., 109 (1992), pp. 76–97.

[AMT] L. AMBROSIO, S. MORTOLA, AND V. M. TORTORELLI, *Functional with linear growth defined on vector-valued BV functions*, J. Math. Pures Appl., 70 (1991), pp. 269–322.

[BFM] G. BOUCHITTÉ, I. FONSECA, AND L. MASCARENHAS, *A global method for relaxation*, Arch. Rational Mech. Anal., to appear.

[C] J. W. CAHN, *Critical point wetting*, J. Chem. Phys., 66 (1977), pp. 3667–3672.

[CH1] J. W. CAHN AND J. E. HILLIARD, *Free energy of a non-uniform system* I: *Interfacial energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[CH2] J. W. CAHN AND J. E. HILLIARD, *Free energy of a nonuniform system.* III. *Nucleation in a two-component incompressible fluid,* J. Chem. Phys., 31 (1959), pp. 688–699.

[DM] G. DAL MASO, *Integral representation on $BV(\Omega)$ of Γ-limits of variational integrals*, Manuscripta Math., 30 (1980), pp. 387–416.

[EG] L. C. EVANS AND R. F. GARIEPY, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.

[F] I. FONSECA, *Lower semicontinuity of surface energies*, Proc. Roy. Soc. Edinburgh Sect. A, 120 (1992), pp. 99–115.

[FM1] I. FONSECA AND S. MÜLLER, *Quasi-convex integrands and lower semicontinuity in L^1*, SIAM J. Math. Anal., 23 (1992), pp. 1081–1098.

[FM2] I. Fonseca and S. Müller, *Relaxation of quasiconvex functionals in $BV(\Omega, \mathbb{R}^p)$ for integrands $f(x, u, \nabla u)$*, Arch. Rational Mech. Anal., 123 (1993), pp. 1–49.

[FR] I. Fonseca and P. Rybka, *Relaxation of multiple integrals in the space $BV(\Omega, \mathbb{R}^p)$*, Proc. Roy. Soc. Edinburgh Sect. A, 121 (1992), pp. 321–348.

[GSe] C. Goffman and J. Serrin, *Sublinear functions of measures and variational integrals*, Duke Math. J., 31 (1964), pp. 159–178.

[GMS] M. Giaquinta, G. Modica, and J. Souček, *Functional with linear growth in the calculus of variations*, Comment. Math. Univ. Carolin., 20 (1974), pp. 143–172.

[G] M. E. Gurtin, *Some results and conjectures in the gradient theory of phase transitions*, in Metastability and Incompletely Posed Problems, IMA Vol. Math. Appl. 3, Springer-Verlag, New York, 1987, pp. 135–146.

[K] M. A. Krasnosel'skii, *Topological Methods in the Theory of Nonlinear Integral Equations*, Pergamon Press, Elmsford, NY, 1964.

[LM] S. Luckhaus and L. Modica, *The Gibbs-Thompson relation within the gradient theory of phase transitions*, Arch. Rational Mech. Anal., 107 (1989), pp. 71–83.

[MP] U. Massari and L. Pepe, *Su di una impostazione parametrica del problema dei capillari*, Ann. Univ. Ferrara, XX (1975), pp. 21–31.

[Mo1] L. Modica, *Gradient theory of phase transitions and minimal interface criterion*, Arch. Rational Mech. Anal., 98 (1987), pp. 123–142.

[Mo2] L. Modica, *Gradient theory of phase transitions with boundary contact energy*, Ann. Inst. H. Poincaré Anal. NonLinèaire, 4 (1987), pp. 485–512.

[M] S. Müller, *On quasiconvex functions which are homogeneous of degree* 1, Indiana Univ. Math. J., 41 (1992), pp. 295–301.

[Re] Y. G. Reshetnyak, *Weak convergence of completely additive vector functions on a set*, Siberian Math. J., 9 (1968), pp. 1386–1394 (in English); Sibirsk. Mat. Žh., 9 (1968), pp. 1386–1394 (in Russian).

[Se1] J. Serrin, *A new definition of the integral for non-parametric problems in the calculus of variations*, Acta Math., 102 (1959), pp. 23–32.

[Se2] J. Serrin, *On the definition and properties of certain variational integrals*, Trans. Amer. Math. Soc., 161 (1961), pp. 139–167.

[V1] A. Visintin, *Models of Phase Transitions*, Birkhäuser-Verlag, Basel, 1996.

[V2] A. Visintin, *Nonconvex functionals related to multiphase systems*, SIAM J. Math. Anal., 21 (1990), pp. 1281–1304.

[vdW] J. D. van der Waals, *The thermodynamic theory of capillarity under the hypothesis of a continuous variation of density*, in Dutch Verhandel. Konink. Akad. Weten. Amsterdam (Section 1), 1 (1893) (in Dutch); J. Statist. Phys., 20 (1979), pp. 197–244 (in English).

[Z] W. Ziemer, *Weakly Differential Functions*, Springer-Verlag, Berlin, New York, 1989.

SIAM J. MATH. ANAL.
Vol. 30, No. 1, pp. 220–232

STABLE DETERMINATION OF BOUNDARIES FROM CAUCHY DATA*

E. BERETTA† AND S. VESSELLA‡

Abstract. The problem of determining a portion Γ of the boundary of a bounded planar domain Ω from Cauchy data arises in several contexts, for example, such as in corrosion detection from electrostatic measurements. We investigate this severely ill-posed problem establishing logarithmic continuous dependence of Γ from Cauchy data.

Key words. inverse problems, stability

AMS subject classifications. 35R25, 35R30

PII. S0036141097325733

1. Introduction. Let Ω be a bounded open subset of $\mathbf{R}^2$ and let Γ indicate a portion of the boundary of Ω. Let ψ be a function on $\partial\Omega$ having zero mean value and such that supp $(\psi)\gamma$, where $\gamma \subset \partial\Omega\backslash\Gamma$. Consider the Neumann problem

$$\begin{cases} \Delta u = 0 & \text{in } \Omega, \\ \frac{\partial u}{\partial n} = 0 & \text{on } \Gamma, \\ \frac{\partial u}{\partial n} = \psi & \text{on } \partial\Omega\backslash\Gamma, \end{cases} \tag{1.1}$$

where n denotes the outer unit normal to $\partial\Omega$. If we normalize u setting, for example,

$$\int_{\partial\Omega} u ds = 0, \tag{1.2}$$

then the solution u to (1.1) is uniquely determined.

In this paper we shall deal with the inverse problem of determining Γ from the knowledge of the additional boundary measurement

$$u|_\gamma = g$$

when one suitable Neumann datum ψ in (1.1) is assigned. This problem occurs, for example, in corrosion detection by electrostatic measurements [KS], [KSV]. In this case Γ represents the damaged part of the boundary of a conducting specimen and one tries to identify Γ by assigning a suitable flux ψ and measuring the corresponding potential u on the accessible part of the boundary, γ.

Another application of this problem is planar crack detection in nonferrous metals from electromagnetic measurements [McI], [AP], [ABJ].

In this paper we address the question of continuous dependence of Γ from the Cauchy data (ψ, g). Since it is necessary to determine the interior values of u from the Cauchy data, one expects the problem to be severely ill-posed. In fact, in [A1] Alessandrini proves that logarithmic type continuous dependence of Γ from the Cauchy data

*Received by the editors July 28, 1997; accepted for publication February 9, 1998; published electronically November 5, 1998.
http://www.siam.org/journals/sima/30-1/32573.html
†Istituto di Analisi Globale e Applicazioni, Via S. Marta 13/A Firenze, Italy (beretta@vaxiac.iac.rm.cnr.it).
‡Dipartimento di Matematica, DIMADEFAS, Via Lombroso 6/17, Firenze, Italy (vessella@stat.ds.unifi.it).

(ψ, g) is the best possible. An indication of the ill-posedness of the problem is also given in [AP].

In this paper we prove that, under suitable a priori assumptions on the unknown curve Γ, the continuous dependence of Γ from the data is in fact of logarithmic type.

In order to give an idea of the proof of our result let us begin by illustrating the proof of the uniqueness of Γ. Let $\psi \neq 0$ and let Γ_1, Γ_2 be two simple curves having common endpoints. Let $u_1 = u_2$ on γ (here $u_j, j = 1, 2$, is harmonic in Ω_j, the bounded smooth subset corresponding to Γ_j). Let Λ be the connected component of $\overline{\Omega}_1 \cap \overline{\Omega}_2$ containing γ. Uniqueness of the Cauchy problem for the Laplace equation and analytic continuation of harmonic functions give that $u_1 = u_2$ in Λ. Therefore, if for instance $\Omega_1 \backslash \Omega_2 \neq \emptyset$, then by the inclusion $\partial(\Omega_1 \backslash \Lambda) \subset (\partial\Lambda \cap \Gamma_2) \cup \Gamma_1$, by the fact that $\frac{\partial u_2}{\partial n} = 0$ on Γ_2, and by the divergence theorem one has

$$\int_{\Omega_1 \backslash \Omega_2} |\nabla u_1|^2 dx \leq \int_{\Omega_1 \backslash \Lambda} |\nabla u_1|^2 dx = \int_{\partial\Lambda \cap \Gamma_2} u_1 \frac{\partial(u_1 - u_2)}{\partial n} ds. \tag{1.3}$$

Therefore, from (1.3) it follows that $|\nabla u_1| = 0$ in $\Omega_1 \backslash \Omega_2$ and by analytic continuation of harmonic functions it follows that $\nabla u_1 = 0$ in Ω_1, contradicting the assumption $\frac{\partial u_1}{\partial n}|_\gamma \neq 0$.

To prove our result we use (1.3) in order to estimate the Lebesgue measure of the set $\Omega_1 \triangle \Omega_2$. In fact, choosing a suitable function ψ, by [A2], it follows that u_1 has no critical points in $\overline{\Omega_1} \backslash \gamma$ and a pointwise lower bound on $|\nabla u_1|$ holds; cf. Lemma 3.3. Coupling this fact with stability estimates for the Cauchy problem applied to $u_1 - u_2$, using (1.3) we derive a first rough stability estimate for the Lebesgue measure of the set $\Omega_1 \triangle \Omega_2$. This estimate is refined in the second part of Theorem 2.1. Finally we obtain the desired estimate on the Hausdorff distance of Γ_1 from Γ_2 establishing an upper bound for the Hausdorff distance of Γ_1 from Γ_2 in terms of the Lebesgue measure of the set $\Omega_1 \triangle \Omega_2$, Proposition 3.4.

The paper is organized as follows. In section 2 we describe the a priori assumptions on Ω, Γ, and on the measurements and we state our main result (Theorem 2.1). In section 3 we derive some bounds on u and on the measure of Ω and of $\partial\Omega$, which are consequences of the a priori information illustrated in section 2. In section 4 we prove Theorem 2.1. Finally in the Appendix we give the proof of Proposition 3.4 of section 3.

2. The main result. We start by giving some notation and by listing the required a priori assumptions.

Let $x^0 \in \mathbf{R}^2$, $r > 0$, A, B be two measurable subsets of $\mathbf{R}^2$. In the sequel we will use the following notation:

(i) $B(x^0, r) = \{x \in \mathbf{R}^2 : |x - x^0| < r\}$;

(ii) $|A|$ denotes the measure of the set A; if ∂A is smooth, then $|\partial A|$ denotes the length of ∂A;

(iii) $d(x^0, A) = \inf_{x \in A} |x - x^0|$, $\quad d(A, B) = \inf_{x \in A} d(x, B)$;

(iv) $(A)_r = \{x \in A : d(x, \mathbf{R}^2 \backslash A) \geq r\}$;

(v) $\delta_{\mathcal{H}}(A, B) = \sup\{\sup_{x \in A} d(x, B), \sup_{x \in B} d(x, A)\}$, Hausdorff distance between the sets A and B;

(vi) $C = C(a, b, c, \ldots)$ indicates that C depends only on the parameters $a, b, c, \ldots$.

In the sequel we will indicate different constants with the same letter C.

Given the positive constants A, L_0, L_1, R we suppose that Ω indicates a simply connected bounded open C^2 subset of $\mathbf{R}^2$ satisfying the following conditions.

Prior information on Ω.

$$|\Omega| \leq A; \tag{2.1a}$$

$$\begin{cases} \forall x \in \partial\Omega \text{ there exist two disks of radius } R \text{ tangent in } x, \\ \text{the first contained in } \overline{\Omega} \text{ and the second in } \mathbf{R}^2\backslash\Omega. \end{cases} \tag{2.1b}$$

Prior information on the unknown boundary Γ. Let $\gamma \subset \partial\Omega$ denote an arc of endpoints a_0 and a_1 where we will make our measurements. Denote with Γ an arc with $\overline{\Gamma} \subset \text{int}(\partial\Omega\backslash\gamma)$ and of endpoints b_0 and b_1 such that

$$d(\gamma, \Gamma) \geq L_0 \tag{2.2a}$$

and

$$|\Gamma| \geq L_0. \tag{2.2b}$$

Prior information on the measurements. Given two points $p_1, p_2 \in \gamma$ such that

$$|p_1 - p_2| \geq L_1, \tag{2.3}$$

we consider two functions $\eta_1, \eta_2 \in C^2(\partial\Omega)$ such that

$$\int_{\partial\Omega} \eta_i ds = 1, \quad i = 1, 2, \tag{2.4a}$$

and

$$\text{supp}(\eta_i) \subset \gamma \cap B(p_i, h), \quad i = 1, 2, \tag{2.4b}$$

where $h \in (0, L_1/2)$.

Let

$$\psi = \eta_1 - \eta_2. \tag{2.5}$$

We will consider the solution u of the following Neumann problem

$$\begin{aligned} \Delta u &= 0 && \text{in } \Omega, \\ \tfrac{\partial u}{\partial n} &= 0 && \text{on } \Gamma, \\ \tfrac{\partial u}{\partial n} &= \psi && \text{on } \partial\Omega\backslash\Gamma, \\ \textstyle\int_{\partial\Omega} u ds &= 0, \end{aligned} \tag{2.6}$$

where n denotes the exterior unit normal to $\partial\Omega$. Observe that by (2.4a)–(2.4b) $\psi = 0$ on $\partial\Omega\backslash\gamma$.

Given two curves Γ_i, $i = 1, 2$, let u_1, u_2 indicate the solutions of problem (2.6) corresponding to the curves Γ_1, Γ_2, respectively. Then the following result holds.

THEOREM 2.1. *Assume conditions* (2.1)–(2.5). *Given* $\epsilon \in (0, 1)$, *if*

$$\max_{x\in\gamma} |u_1(x) - u_2(x)| \leq \epsilon,$$

then for any $\eta > 0$ *there exists a positive constant* $C = C(L_0, L_1, A, R, \eta)$ *such that*

$$\delta_{\mathcal{H}}(\Gamma_1, \Gamma_2) \leq C |\ln \epsilon|^{-\frac{1}{3}+\eta}.$$

See section 4 for the proof.

3. Some preliminary results. In this section we will expose and prove some consequences of the a priori information listed in the previous section.

PROPOSITION 3.1. *Let the a priori conditions* (2.1a) *and* (2.1b) *be satisfied. Let* $\kappa(x)$ *indicate the curvature at the point* $x \in \partial\Omega$. *Then, for any* $r \in [0, R]$, *the following bounds hold:*

$$|\kappa(x)| \leq \frac{1}{R} \tag{3.1}$$

for any $x \in \partial\Omega$. *We also have that*

$$|\partial\Omega| \leq c\frac{A}{R}, \tag{3.2}$$

$$|\Omega\backslash(\Omega)_r| \leq c\frac{A}{R}r, \tag{3.3}$$

and

$$|\partial(\Omega)_r| \leq c\frac{A}{R}, \tag{3.4}$$

where $c = 24\sqrt[3]{6}e^2$.

Proof. The bound (3.1) follows immediately from assumption (2.1b).

By [V] one has that for $r \in [0, R]$,

$$|\partial(\Omega)_r| = |\partial\Omega| - r\int_{\partial\Omega} \kappa(x)ds \tag{3.5}$$

and

$$|\Omega\backslash(\Omega)_r| = \int_0^r |\partial(\Omega)_t|dt = r|\partial\Omega| - \frac{r^2}{2}\int_{\partial\Omega} \kappa(x)ds. \tag{3.6}$$

Let $f(r) = |\Omega\backslash(\Omega)_r|$ for $r \in [0, R]$. By [G] the interpolation inequality for f follows:

$$\begin{aligned} |\partial\Omega| = f'(0) &\leq 12e^2\|f\|_\infty^{2/3}\left(\|f'''\|_\infty + \frac{6}{R^2}\|f\|_\infty\right)^{1/3} \\ &\leq 12\sqrt[3]{6}e^2\frac{|\Omega|}{R} \leq 12\sqrt[3]{6}e^2\frac{A}{R}, \end{aligned} \tag{3.7}$$

which gives estimate (3.2). From (3.5), (3.1), and (3.7) it follows that for $r \in [0, R]$

$$|\partial(\Omega)_r| \leq |\partial\Omega| + \frac{r}{R}|\partial\Omega| \leq 2|\partial\Omega| \leq 24\sqrt[3]{6}e^2\frac{A}{R}$$

from which one gets (3.4). Finally (3.6), (3.1), and (3.7) lead to

$$|\Omega\backslash(\Omega)_r| \leq r|\partial\Omega|\left(1 + \frac{r}{2R}\right) \leq 18\sqrt[3]{6}e^2\frac{A}{R}$$

for $r \in [0, R]$ and (3.3) follows, completing the proof. □

Observe that from bound (3.2) of Proposition 3.1 it follows immediately that

$$diam(\Omega) \leq c\frac{A}{R}. \tag{3.8}$$

Now we derive some upper and lower bounds for the solution u of problem (2.6).

PROPOSITION 3.2. *Let the a priori assumptions* (2.1a)–(2.1b) *be satisfied. Then for any $q > 2$ there exist three positive constants $E_i = E_i(A, R, \|\psi\|_{C^2(\partial\Omega)}, q)$, $i = 0, 1, 2$, such that*

$$\|u\|_{L^\infty(\Omega)} \le E_0, \quad \|\nabla u\|_{L^\infty(\Omega)} \le E_1, \quad \|D^2 u\|_{L^q(\Omega)} \le E_2. \tag{3.9}$$

Proof. Let $\phi(s) = \int_0^s \psi(t)dt$, where s is the arclength parameter on $\partial\Omega$. Then the harmonic conjugate v, of u, satisfies the Dirichlet problem

$$\begin{aligned} \Delta v &= 0 \qquad \text{in } \Omega, \\ v &= \phi \qquad \text{on } \partial\Omega. \end{aligned} \tag{3.10}$$

Let $\Phi \in C^2(\overline{\Omega})$ be an extension of ϕ to $\overline{\Omega}$ such that $\Phi|_{\partial\Omega} = \phi$ and $\|\Phi\|_{C^2(\overline{\Omega})} \le C\|\phi\|_{C^2(\partial\Omega)}$. Then a priori estimates for elliptic equations [GT] give

$$\|v\|_{2,q} \le C\|\psi\|_{C^2(\partial\Omega)}.$$

From Sobolev immersion theorem and by the fact that

$$|\nabla u(x)| = |\nabla v(x)|, \quad |D^2 u(x)| = |D^2 v(x)|,$$

it follows that

$$\|u\|_{L^\infty(\Omega)} \le E_0, \quad \|\nabla u\|_{L^\infty(\Omega)} \le E_1, \quad \|D^2 u\|_{L^q(\Omega)} \le E_2,$$

where E_0, E_1, E_2 depend on $A, R, \|\psi\|_{C^2(\partial\Omega)}$, and q, which then proves the proposition. □

LEMMA 3.3. *Assume conditions* (2.1)–(2.5) *hold and let u be the solution of problem* (2.6). *Then there exists a positive constant $C = C(L_0, A, R, \|\psi\|_{C^2(\partial\Omega)})$ such that*

$$|\nabla u(x)| \ge C \quad \textit{for any } x \in \Omega : d(x, \gamma) \ge \frac{L_0}{2}. \tag{3.11}$$

Proof. As in Proposition 3.2 we consider the harmonic conjugate of u, v satisfying

$$\begin{aligned} \Delta v &= 0 \qquad \text{in } \Omega, \\ v &= \phi \qquad \text{on } \partial\Omega, \end{aligned}$$

where $\phi = \int_0^s \psi(t)dt$. Let $z = x + iy$ and let $f(z) = \xi + i\eta = \zeta$ denote a conformal mapping such that

(a) f maps Ω one to one onto $B^+(0,1) = \{\zeta : |\zeta| < 1, \eta > 0\}$;
(b) $f(\gamma)$ consists of the semicircle $\{\zeta : |\zeta| = 1, \eta > 0\}$ and $f(\partial\Omega\backslash\gamma)$ consists of the segment $\{\zeta : \xi \in [-1, 1], \eta = 0\}$.

Then $f \in C^2(\overline{\Omega}\backslash\{a_0, a_1\})$ and the following bounds for $|f'|$ hold:

$$k_1^{-1}\frac{R}{A} \le |f'| \le k_2\frac{1}{L_0} \quad \text{in } \overline{\Omega}, \tag{3.12}$$

where $k_1 = (12\sqrt{2}\sqrt[3]{6}e^2)$ and $k_2 = \sqrt{2}\pi$.

The regularity of f follows from the fact that $\partial\Omega \in C^2$ and $f|_{\partial\Omega}$ is smooth except at the endpoints of γ, a_0, and a_1. To prove the lower bound in (3.12) let us consider

the inverse mapping $f^{-1} : B^+(0,1) \to \Omega$. Then $f^{-1} \in C^2(\overline{B^+(0,1)}\backslash\{q_0, q_1\})$, where q_0, q_1 are the singular points of $\partial B^+(0,1)$. Moreover, if $f^{-1}(\zeta) = \varphi + i\omega$, then by the Cauchy–Riemann equations and by estimate (3.2) of Proposition 3.1 it is easy to see that

$$|\varphi_\xi|, |\varphi_\eta| \le k_1 \frac{A}{R} \text{ on } \partial B^+(0,1). \tag{3.13}$$

By the weak maximum principle it follows that (3.13) holds in $\overline{B^+(0,1)}$. Hence

$$|(f^{-1})'| = |\nabla\varphi| \le k_1 \frac{A}{R} \quad \text{in } \overline{B^+(0,1)}$$

from which the lower bound in (3.12) follows. To show the upper bound for $|f'|$ we proceed similarly estimating $|\xi_x|, |\xi_y|$ on $\partial\Omega$ and using the a priori information (2.2b).

Let $w = v \circ f^{-1} : B^+(0,1) \to \mathbf{R}$ and consider the odd reflection of w, w^* satisfying

$$\begin{aligned} \Delta w^* &= 0 \qquad \text{in } B(0,1), \\ w^* &= \phi^* \qquad \text{on } \partial B(0,1). \end{aligned}$$

By our choice of ϕ, $\partial B(0,1)$ can be split into two arcs on which alternatively ϕ^* is a nondecreasing and nonincreasing function of the arclength parameter. Then, by Theorem 2.7 of [AM], w^* has no interior critical points. Furthermore by [A2] there exists a positive constant $C = C(d, \|\phi^*\|_{C^2(\overline{B(0,1)})})$ such that

$$|\nabla w^*(\zeta)| \ge C \quad \text{for any } \zeta \in B(0,1) : d(\zeta, \partial B(0,1)) \ge d > 0.$$

Hence

$$|\nabla w(\zeta)| \ge C \quad \text{for any } \zeta \in B^+(0,1) : d(\zeta, \partial B^+(0,1)\backslash\{\eta = 0\}) \ge d > 0. \tag{3.14}$$

Let $\zeta \in B^+(0,1)$ satisfy (3.14) and let $z \in \Omega$ be such that $\zeta = f(z)$. Then $d(z,\gamma) \ge \frac{dL_0}{k_1}$. Choose $d = \frac{k_1}{4}$. Then by (3.13) and (3.14) it follows that

$$|\nabla u(z)| = |\nabla v(z)| = |\nabla w(f(z))| \cdot |f'(z)| \ge C > 0$$

for $z \in \Omega$ such that $d(z,\gamma) \ge \frac{L_0}{4}$. Finally, since $\phi \circ f^{-1} \in C^2(\overline{B^+(0,1)})$, ψ has support in γ and $f^{-1} \in C^2(\overline{B^+(0,1)}\backslash\{q_0, q_1\})$ one has that $C = C(L_0, A, R, \|\psi\|_{C^2(\partial\Omega)})$, which completes the proof. □

PROPOSITION 3.4. *Let Ω_1, Ω_2 denote two bounded simply connected C^2 subsets of $\mathbf{R}^2$ satisfying conditions* (2.1a)–(2.1b) *and such that $\Omega_1, \Omega_2 \subset B(z, 2d)$, where $d = \max\{diam(\Omega_1), diam(\Omega_2)\}$ and $z \in \mathbf{R}^2$. Then*

$$\delta_{\mathcal{H}}(\partial\Omega_1, \partial\Omega_2) \le \frac{16d}{(\pi)^{1/3} R^{2/3}} |\Omega_1 \triangle \Omega_2|^{1/3}. \tag{3.15}$$

Since the proof is rather technical we postpone the proof to the Appendix.

4. Proof of Theorem 2.1. Let u_i be the solution of the problem

$$\begin{aligned} \Delta u_i &= 0 \qquad \text{in } \Omega_i, \\ \frac{\partial u_i}{\partial n} &= 0 \qquad \text{on } \Gamma_i, \\ \frac{\partial u_i}{\partial n} &= \psi \qquad \text{on } \partial\Omega_i\backslash\Gamma_i, \\ \int_{\partial\Omega_i} u_i ds &= 0, \end{aligned} \tag{4.1}$$

where Ω_i indicates the subset corresponding to Γ_i.

Proof of Theorem 2.1. Every time we will say in the sequel that a property is true for ϵ sufficiently small, we will mean that there exists a positive number $\epsilon_0 = \epsilon_0(A, R, L_0, L_1, \|\psi\|_{C^2(\partial\Omega)}, q)$ such that for every $\epsilon \in (0, \epsilon_0)$ the property holds.

Let $R_1 = \min(R, L_0)$ and $u = u_1 - u_2$ be defined in $\Omega_1 \cap \Omega_2$. Let

$$J_1 = \left\{ x \in \Omega_1 \cap \Omega_2 | d(x, \gamma) \leq R_1; d(x, a_0), d(x, a_1) \geq \frac{R_1}{4} \right\}$$

and

$$J_2 = \left\{ x \in \Omega_1 \cap \Omega_2 | d(x, \gamma) \leq R_1; d(x, a_0), d(x, a_1) \geq \frac{R_1}{3} \right\}.$$

Then by [P] it follows that for $x \in J_1$,

$$|\nabla u(x)| \leq C\epsilon^{\lambda}, \tag{4.2}$$

where $\lambda \in (0, 1)$ depends on γ and R_1.

We divide the proof into two steps.

Step 1. In this first part of the proof we establish a first rough estimate of $\delta_{\mathcal{H}}(\Gamma_1, \Gamma_2)$. For let $r \in (0, R_1/3)$. In the sequel we will indicate with γ_r the set $\gamma_r = \{x \in \Omega_1 | d(x, \gamma) = r\}$ and with Λ_r the connected component of $(\Omega_1)_r \cap (\Omega_2)_r$ containing γ_r. The following inclusions hold:

$$\Omega_1 \backslash \Omega_2 \subset [\Omega_1 \backslash (\Omega_1)_r] \cup [(\Omega_1)_r \backslash (\Omega_2)_r], \tag{4.3}$$

$$(\Omega_1)_r \backslash (\Omega_2)_r \subset (\Omega_1)_r \backslash \Lambda_r. \tag{4.3i}$$

Setting $\Gamma'_r = \partial\Omega_j \backslash \gamma$ for $j = 1, 2$, we have that the set $\Gamma'_{j,r}$ indicates the set $\Gamma'_{i,r} = \{x \in \Omega_i | d(x, \Gamma'_i) = r\}$:

$$\partial[(\Omega_1)_r \backslash \Lambda_r] \subset [\partial\Lambda_r \cap \Gamma'_{2,r}] \cup \Gamma'_{1,r}. \tag{4.3ii}$$

From (3.9), (4.3), and (4.3ii) it follows that

$$\int_{\Omega_1 \backslash \Omega_2} |\nabla u_1|^2 dx \leq E_1^2 |\Omega_1 \backslash (\Omega_1)_r| + \int_{(\Omega_1)_r \backslash \Lambda_r} |\nabla u_1|^2 dx. \tag{4.4}$$

From the divergence theorem, from (3.9), and from (4.3ii) one has

$$\begin{aligned} \int_{(\Omega_1)_r \backslash \Lambda_r} |\nabla u_1|^2 dx &= \int_{\partial[(\Omega_1)_r \backslash \Lambda_r]} u_1 \frac{\partial u_1}{\partial n} ds \\ &\leq E_0 \left(\int_{\partial\Lambda_r \cap \Gamma'_{2,r}} \left| \frac{\partial u_1}{\partial n} \right| ds + \int_{\Gamma'_{1,r}} \left| \frac{\partial u_1}{\partial n} \right| ds \right). \end{aligned} \tag{4.5}$$

Now, let $x \in \partial\Lambda_r \cap \Gamma'_{2,r}$ and let $y \in \Gamma'_2$ be such that $x = y - rn_y$, where n_y indicates the exterior outer unit normal at $\partial\Omega_2$ at the point y. Since $n_x = n_y$ and $\frac{\partial u_2}{\partial n} = 0$ on Γ'_2 it follows that

$$\left| \frac{\partial u_1}{\partial n}(x) \right| \leq |\nabla u(x)| + |\nabla u_2(x) - \nabla u_2(y)|. \tag{4.6}$$

From (3.9), from the Sobolev immersion theorem, and from (4.6) one derives

$$\left|\frac{\partial u_1}{\partial n}(x)\right| \le |\nabla u(x)| + Cr^{1-2/q} \text{ for } x \in \partial\Lambda_r \cap \Gamma'_{2,r} \text{ and } q > 2. \tag{4.7}$$

Furthermore by (3.9) and since $\frac{\partial u_1}{\partial n} = 0$ on Γ'_1 one has

$$\left|\frac{\partial u_1}{\partial n}(x)\right| \le Cr^{1-2/q} \text{ for } x \in \Gamma'_{1,r}. \tag{4.8}$$

Hence by (4.4)–(4.8) and by the a priori information we derive

$$\begin{aligned}\int_{\Omega_1\setminus\Omega_2} |\nabla u_1|^2 dx &\le E_1^2 |\Omega_1 \setminus (\Omega_1)_r| \\ &+ E_0 \left((|\Gamma'_{1,r}| + |\Gamma'_{2,r}|) Cr^{1-2/q} + |\Gamma'_{2,r}| \max_{\partial\Lambda_r\cap\Gamma'_{2,r}} |\nabla u| \right) \\ &\le C \left[\left(\frac{r}{R_1}\right)^{1-2/q} + \max_{\partial\Lambda_r\cap\Gamma'_{2,r}} |\nabla u| \right].\end{aligned} \tag{4.9}$$

We now estimate $\max_{\partial\Lambda_r\cap\Gamma'_{2,r}} |\nabla u|$. More precisely we will prove that

$$|\nabla u(x)| \le C(1+\epsilon^\lambda)^{1-\alpha^{N_r}} \epsilon^{\lambda\alpha^{N_r}} \text{ for } x \in \partial\Lambda_r \cap \Gamma'_{2,r}, \tag{4.10}$$

where $\alpha = \frac{\ln 4/3}{\ln 4}$, $N_r = \frac{A}{\pi(\frac{r}{4})^2} + 1$, and $\lambda \in (0,1)$.

Now let $\bar{x} \in \partial\Lambda_r \cap \Gamma'_{2,r}$ and $y \in \gamma_r \cap J_2$. Then one has that $B(y, r/4) \subset J_1$ for $r < R_1/3$. Set

$$\sigma = \max_{\overline{B(y,r/4)}} |\nabla u|. \tag{4.11}$$

Let $\mathcal{L}$ be a simple curve contained in Λ_r having $\bar{x}$ and y as endpoints. Then it is possible to construct a chain of closed balls centered in $\mathcal{L}$, of radius $r/4$, tangent two by two, and internally nonoverlapping; the first ball is centered in y; the last is at distance less than or equal to $r/4$ from $\bar{x}$. If the distance is less than $r/4$, we add to the chain the ball of radius $r/4$ whose center is $\bar{x}$. The repeated use of the three circles theorem leads to

$$\begin{aligned}\|\nabla u\|_{L^\infty(B(x_i,r/4))} &\le \|\nabla u\|_{L^\infty(B(x_{i-1},3r/4))} \\ &\le \|\nabla u\|^{1-\alpha}_{L^\infty(B(x_{i-1},r))} \cdot \|\nabla u\|^{\alpha}_{L^\infty(B(x_{i-1},r/4))}, \ i = 1,2,\dots,n,\end{aligned}$$

where $\alpha = \frac{\ln 4/3}{\ln 4}$. Hence from (4.11) we derive

$$|\nabla u(\bar{x})| \le (2E_1 + \sigma)^{1-\alpha^{n+1}} \sigma^{\alpha^{n+1}}.$$

On the other hand

$$n \le \frac{|\Lambda|}{\pi(\frac{r}{4})^2} \le \frac{A}{\pi(\frac{r}{4})^2},$$

where Λ is the connected component of $\overline{\Omega}_1 \cap \overline{\Omega}_2$ containing γ. This last bound and (4.2) give (4.10). From (4.9) and (4.10) it follows that for $r \in (0, R_1/3)$,

$$\int_{\Omega_1 \setminus \Omega_2} |\nabla u_1|^2 dx \leq C \left[\left(\frac{r}{R_1} \right)^{1-2/q} + (1+\epsilon^\lambda)^{1-\alpha^{N_r}} \cdot \epsilon^{\lambda \alpha^{N_r}} \right]. \tag{4.12}$$

Let $r = \frac{R_1}{4} \left[\frac{C}{\ln |\ln \epsilon^\lambda|} \right]^{1/2}$. Then from (4.12) for ϵ sufficiently small one has

$$\int_{\Omega_1 \setminus \Omega_2} |\nabla u_1|^2 dx \leq C \left(\ln |\ln \epsilon^\lambda| \right)^{-\frac{1}{2}(1-\frac{2}{q})},$$

and by (3.11) one gets

$$|\Omega_1 \setminus \Omega_2| \leq C \left(\ln |\ln \epsilon^\lambda| \right)^{-\frac{1}{2}(1-\frac{2}{q})}.$$

Since a similar bound holds for $|\Omega_2 \setminus \Omega_1|$, it follows that

$$|\Omega_1 \triangle \Omega_2| \leq C \left(\ln |\ln \epsilon^\lambda| \right)^{-\frac{1}{2}(1-\frac{2}{q})}. \tag{4.13}$$

Finally, (4.13) and (3.15) lead to

$$\delta_{\mathcal{H}}(\partial\Omega_1, \partial\Omega_2) \leq C \left(\ln |\ln \epsilon^\lambda| \right)^{-\frac{1}{6}(1-\frac{2}{q})} \tag{4.14}$$

for ϵ sufficiently small.

Step 2. From the divergence theorem and from (3.2) it follows that

$$\int_{\Omega_1 \setminus \Omega_2} |\nabla u_1|^2 dx \leq \frac{cE_0 A}{R} \cdot \sup_{(\partial\Lambda \cap \Gamma_2') \setminus \Gamma_1'} |\nabla u|. \tag{4.15}$$

We will assume that ϵ is sufficiently small in such a way that

$$C \left(\ln |\ln \epsilon^\lambda| \right)^{-\frac{1}{6}(1-\frac{2}{q})} \leq \frac{R_1}{2}. \tag{4.16}$$

Let $\bar{x} \in (\partial\Lambda \cap \Gamma_2') \setminus \Gamma_1'$ and let $r = d(\bar{x}, \partial\Omega_1)$. From (4.14) and (4.16) it follows that $\bar{x} \in \partial(\Omega_1)_r$ and, moreover, $(\Omega_1)_r$ has the interior sphere property with radius $R - r$ and the exterior sphere property with radius R.

Let $B_{int}(p, R_1/2)$ and $B_{ext}(q, R_1/2)$ indicate the interior and exterior disks to $(\Omega_1)_r$ tangent in $\bar{x}$ to $\partial(\Omega_1)_r$. Let $B_{int}(p', R_1/2)$ and $B_{ext}(q', R_1/2)$ be the interior and exterior disks to Ω_2 tangent in $\bar{x}$ to $\partial\Omega_2$. We have that

$$B_{int}(p', R_1/2) \cap B_{ext}(q, R_1/2) \subset \Omega_2 \setminus (\Omega_1)_r. \tag{4.17}$$

From (3.3), (4.13), (4.16), and (4.17) one has

$$B_{int}(p', R_1/2) \cap B_{ext}(q, R_1/2) \leq C \left(\ln |\ln \epsilon^\lambda| \right)^{-\frac{1}{6}(1-\frac{2}{q})}. \tag{4.18}$$

Let $\bar{y}$ be the second intersection of $\partial B_{int}(p, R_1/2)$ with $\partial B_{int}(p', R_1/2)$ and let δ denote the angle $\bar{y}\hat{\bar{x}}p'$. Since

$$\left| B_{int}\left(p', \frac{R_1}{2}\right) \cap B_{ext}\left(q, \frac{R_1}{2}\right) \right| = \frac{R_1^2}{4} \left(2\delta - \sin 2\delta\right),$$

by (4.18) it follows that

$$\delta \leq 2\delta(\epsilon),$$

where $\delta(\epsilon)$ is defined by the relation

$$\frac{R_1^2}{4}\left(2(2\delta(\epsilon)) - \sin 2(2\delta(\epsilon))\right) = C\left(\ln|\ln\epsilon^{\lambda}|\right)^{-\frac{1}{6}(1-\frac{2}{q})}.$$

Denote with $B(m,\rho)$ the disk centered at $m = \frac{p+p'}{2}$ of radius $\rho = 1 - \sin\delta$. Then it is easy to see that $B(m,\rho)$ is the maximum disk contained in $B_{int}(p, R_1/2) \cap B_{int}(p', R_1/2)$.

Let $\delta_0 \in (0, \pi/12)$ be fixed and let ϵ be sufficiently small so that

$$\delta(\epsilon) \leq \delta_0^2.$$

Let us describe the following geometric construction: consider the half-line originating from $\bar{x}$ intersecting $\partial B_{int}(p, R_1/2)$ at s and forming an angle of width $\pi/2 - \theta$, $\theta \in (4\delta_0, \pi/2)$, with the half-line $\bar{x}m$. Now let $m' = \frac{\bar{x}+s}{2}$. As θ varies in the interval $(4\delta_0, \pi/2)$, the point m' describes a curve of endpoints a and m. Now reflect this curve across the line $\bar{x}m$ and denote with β the entire curve. From the three circles theorem, one has

$$|\nabla u(x)| \leq (2E_1 + \sigma)^{1-\omega_0}\,\sigma^{\omega_0} \quad \forall x \in \beta,$$

where

$$\omega_0 = \frac{\ln\left(\frac{\cos 3\delta_0}{\cos 4\delta_0}\right)}{\ln 2}$$

and

$$\sigma = \max_{B(m,\rho/2)} |\nabla u|.$$

A Carleman estimate in a sector [C] leads to

$$|\nabla u(q)| \leq (2E_1 + \sigma)\left(\frac{\sigma}{2E_1+\sigma}\right)^{\omega_0\left(\frac{l}{R_1/2\sin 2\delta_0}\right)^{\frac{1}{1-\frac{8\delta_0}{\pi}}}},$$

where q lies on the segment $\bar{x}m$ at distance l from $\bar{x}$.

Setting $t = \frac{l}{\frac{R_1}{2}\sin 2\delta_0}$ one has that $t \in (0,1)$ and, moreover,

$$|\nabla u(\bar{x})| \leq C\left(\frac{tR_1}{2}\right)^{1-\frac{2}{q}} + (2E_1+\sigma)\left(\frac{\sigma}{2E_1+\sigma}\right)^{\omega_0 t^{\frac{1}{1-\frac{8\delta_0}{\pi}}}}. \tag{4.19}$$

For $\sigma < \frac{2E_1}{e}$ let $t = |\ln\left(\frac{\sigma}{2E_1+\sigma}\right)|^{-(1-\frac{9\delta_0}{\pi})}$. Then from (4.19) for σ small enough it follows that

$$|\nabla u(\bar{x})| \leq C\left|\ln\left(\frac{\sigma}{2E_1+\sigma}\right)\right|^{-(1-\frac{2}{q})(1-\frac{9\delta_0}{\pi})}.$$

Finally let us estimate σ. Using similar arguments as for the proof of estimate (4.10) one has

$$\sigma \leq C(1+\epsilon^{\lambda})^{1-\alpha^{N_0}}\epsilon^{\lambda\alpha^{N_0}},$$

where $\alpha = \frac{\ln 4/3}{\ln 4}$, $N_0 = \frac{2A}{\pi(\frac{R_1}{32})^2} + 1$, and $\lambda \in (0,1)$. Then for ϵ sufficiently small

$$|\nabla u(\bar{x})| \leq C|\ln\epsilon|^{-(1-\frac{2}{q})(1-\frac{9\delta_0}{\pi})}.$$

The last inequality, (4.15), and (3.11) give

$$|\Omega_1\backslash\Omega_2| \leq C|\ln\epsilon|^{-(1-\frac{2}{q})(1-\frac{9\delta_0}{\pi})}.$$

Since an analogous bound holds for $|\Omega_2\backslash\Omega_1|$ we derive

$$|\Omega_1 \triangle \Omega_2| \leq C|\ln\epsilon|^{-(1-\frac{2}{q})(1-\frac{9\delta_0}{\pi})}.$$

Finally Proposition 3.4, last inequality, and the fact that

$$\delta_{\mathcal{H}}(\partial\Omega_1, \partial\Omega_2) = \delta_{\mathcal{H}}(\Gamma_1, \Gamma_2)$$

give

$$\delta_{\mathcal{H}}(\Gamma_1, \Gamma_2) \leq C|\ln\epsilon|^{-\frac{1}{3}+\eta},$$

which proves Theorem 2.1. □

5. Appendix.

Proof of Proposition 3.4. In order to prove inequality (3.15) we prove preliminarily that

$$\delta_{\mathcal{H}}(\Omega_1, \Omega_2) \leq \frac{4d}{(\pi)^{1/3}R^{2/3}}|\Omega_1 \triangle \Omega_2|^{1/3}. \tag{5.1}$$

Let $\sigma = |\Omega_1 \triangle \Omega_2|$. If $\sigma \geq \frac{\pi R^2}{8}$, then

$$\delta_{\mathcal{H}}(\Omega_1, \Omega_2) \leq 2d \leq 2d\left(\frac{\sigma}{\frac{\pi R^2}{8}}\right)^{1/3} = \frac{4d}{R^{2/3}\pi^{1/3}}\sigma^{1/3},$$

which proves (5.1).

Let us consider the case $\sigma < \frac{\pi R^2}{8}$. Let $\bar{x} \in \overline{\Omega}_1\backslash\Omega_2$. From property (2.1b) there exists a point $p \in \Omega_1$ such that $B(p, R/2) \subset \Omega_1$ and $\bar{x} \in \partial B(p, R/2)$. Since $\sigma < \frac{\pi R^2}{8}$ it is easy to see that $\overline{B(p, R/2)} \cap \Omega_2 \neq \emptyset$.

Let $y \in \partial\Omega_2 \cap \overline{B(p, R/2)}$ be such that

$$|\bar{x} - y| = d(\bar{x}, \overline{\Omega}_2 \cap \overline{B(p, R/2)}). \tag{5.2}$$

Again by (2.1b) there exists a point $q \in \mathbf{R}^2\backslash\Omega_2$ such that $B(q, R) \subset \mathbf{R}^2\backslash\Omega_2$ and $\partial B(q, R)$ is tangent to $\partial\Omega_2$ at y. We now prove that

$$\bar{x} \in \overline{B(p, R/2)} \cap \overline{B(q, R)}. \tag{5.3}$$

We distinguish two cases

(i) $y \in B(p, R/2)$,

(ii) $y \in \partial B(p, R/2)$.

In case (i) observe that by construction $(\bar{x} - y) \perp \partial\Omega_2$. Moreover, $|\bar{x} - y| < diam B(p, R/2) = R$ from which (5.3) follows.

Let us now prove (5.3) in case (ii). From simple but lengthy geometric calculations it follows that the outer normal n at $\partial\Omega_2$ in y belongs to the angle $\bar{x}\hat{y}p$. This fact implies that the center q of the disk $B(q, R)$ must lie on the arc $\overset{\frown}{q_1 q_2}$ of the circle centered at y of radius R and with endpoints at q_1 and q_2, where q_1 is the contact point of $\partial B(p, R/2)$ and $\partial B(y, R)$ and q_2 is the intersection of the half-line $y\bar{x}$ with $\partial B(y, R)$. Then one easily verifies

$$|\bar{x} - q| \leq |\bar{x} - q_1| \leq |y - q_1| = R.$$

Therefore $\bar{x} \in \overline{B(q, R)}$ which concludes the proof of (5.3). Set $K = \overline{B(p, R/2)} \cap \overline{B(q, R)}$. Since $|\bar{x} - y| \leq diam K$, the rest of the proof consists of finding a bound for $diam K$ in terms of σ. Without loss of generality we may assume that $q = (0, 0)$ and that $p = (0, a)$ with $a \geq 0$. We observe that a cannot be greater than $3(R/2)$ since $K \neq \emptyset$; on the other hand $a \notin [0, \sqrt{3}R/2]$ since $\sigma < \frac{\pi R^2}{8}$. Let us now consider the case $a \in [R, \frac{3R}{2}]$. (For the case $a \in [\sqrt{3}R/2, R]$ we may proceed analogously observing that $|K| \geq |B(0, R) \cap B((0, R), R/2)|$.) Simple geometric arguments lead to the fact that $diam K = |m_1 - m_2|$, where m_1 and m_2 denote the intersections of $\partial B((0, a), R/2)$ with $\partial B((0, 0), R)$. Indicating with 2θ the angle $m_1 \overset{\wedge}{0} m_2$ one has

$$\sigma \geq |K| \geq R^2(2\theta - \sin 2\theta). \tag{5.4}$$

Finally from (5.4) it is easy to see that

$$\theta \leq \left(\frac{\sigma}{R^2}\right)^{1/3}. \tag{5.5}$$

By (5.5) we have that for $\bar{x} \in \Omega_1$,

$$d(\bar{x}, \Omega_2) = |\bar{x} - y| \leq diam K = |m_1 - m_2| = 2R \sin\theta \leq 2R \left(\frac{\sigma}{R^2}\right)^{1/3}. \tag{5.6}$$

Similarly if $\bar{x} \in \Omega_2$,

$$d(\bar{x}, \Omega_1) \leq 2R \left(\frac{\sigma}{R^2}\right)^{1/3},$$

which gives (5.1).

Let us finally prove (3.15). Let $\Omega_1' = B_{4d} \backslash \Omega_1$ and $\Omega_2' = B_{4d} \backslash \Omega_2$, where B_{4d} is a disk concentric with B_{2d} and radius $4d$.

Since $\Omega_1' \triangle \Omega_2' = \Omega_1 \triangle \Omega_2$ one has

$$\delta_{\mathcal{H}}(\Omega_1', \Omega_2') \leq \frac{16d}{R^{2/3}\pi^{1/3}} \sigma^{1/3} = \beta.$$

We claim that

$$\delta_{\mathcal{H}}(\partial\Omega_1, \partial\Omega_2) \leq \beta.$$

In fact, let $x \in \partial\Omega_1$.

If $x \notin \partial\Omega_2$, then

$$d(x, \partial\Omega_2) = d(x, \Omega_2) \leq \delta_{\mathcal{H}}(\Omega_1, \Omega_2) \leq \beta.$$

If $x \in \partial\Omega_2$, then

$$d(x, \Omega_2') \leq \delta_{\mathcal{H}}(\Omega_1', \Omega_2') \leq \beta. \tag{5.7}$$

Since

$$d(x, \Omega_2') = d(x, \partial\Omega_2)$$

from (5.7) we derive that

$$d(x, \partial\Omega_2) \leq \beta.$$

Analogously if $x \in \partial\Omega_2$, one has

$$d(x, \partial\Omega_1) \leq \beta,$$

which concludes the proof. □

REFERENCES

[ABJ] S. Andrieux, A. Ben Abda, and M. Jaoua, *Identifiabilité de frontière inacessible par des mesures de surface*, C.R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 429–434.

[A1] G. Alessandrini, *Examples of instability in inverse boundary value problems*, Inverse Problems, 13 (1997), pp. 887–897.

[A2] G. Alessandrini, *An identification problem for an elliptic equation in two variables*, Ann. Mat. Pura Appl., 4 (1986), pp. 265–296.

[AM] G. Alessandrini and R. Magnanini, *Elliptic equations in divergence form, geometric critical points of solutions, and Stekloff eigenfunctions*, SIAM J. Math. Anal., 25 (1994), pp. 1259–1268.

[AP] N. D. Aparicio and M. K. Pidcock, *The boundary inverse problem for the Laplace equation in two dimensions*, Inverse Problems, 12 (1996), pp. 565–577.

[C] T. Carleman, *Fonctions quasi-analytiques*, Gauthier–Villars, Paris, 1926, pp. 3–5.

[KS] P. Kaup and F. Santosa, *Nondestructive evaluation of corrosion damage using electrostatic measurements*, J. Nondestructive Evaluation, 14 (1995), pp. 127–136.

[KSV] P. Kaup, F. Santosa, and M. Vogelius, *A method for imaging corrosion damage in thin plates from electrostatic data*, Inverse Problems, 12 (1996), pp. 279–293.

[G] A. Gorny, *Contribution a l'étude des functions derivables d'une variable reelle*, Acta Math., 71 (1939), pp. 317–358.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[McI] M. McIver, *An inverse problem in electromagnetic crack detection*, IMA J. Appl. Math., 47 (1991), pp. 127–145.

[V] S. Vessella, *Stability estimates in an inverse problem for a three-dimensional heat equation*, SIAM J. Math. Anal., 28 (1997), pp. 1354–1370.

[P] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, SIAM, Philadelphia, 1975.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 233–272

MULTIVARIABLE ORTHOGONAL POLYNOMIALS AND COUPLING COEFFICIENTS FOR DISCRETE SERIES REPRESENTATIONS*

HJALMAR ROSENGREN†

Abstract. We study polynomials of several variables which occur as coupling coefficients for the analytic continuation of the holomorphic discrete series of $SU(1,1)$. There are three types of such polynomials, one corresponding to each conjugacy class of one-parameter subgroups. They may be viewed as multivariable generalizations of Hahn, Jacobi, and continuous Hahn polynomials and include many orthogonal and biorthogonal families occurring in the literature. We give a simple and unified approach to these polynomials using the group theoretic interpretation. We prove many formal properties, in particular a number of convolution and linearization formulas. We also develop the corresponding theory for the Heisenberg group, leading to multivariable generalizations of Krawtchouk and Hermite polynomials.

Key words. orthogonal polynomial, biorthogonal polynomials, coupling coefficient, spherical harmonic, multivariable hypergeometric function, convolution formula, linearization formula

AMS subject classifications. 33C50, 33C55, 33C80

PII. S003614109732568X

1. Introduction.

In [R], we introduced three classes of polynomials in several variables. They appeared there as coupling coefficients for certain representations of the group $SU(1,1) \simeq SL(2,\mathbb{R})$ and its covering groups. Special cases occur in the work of many authors, cf. [A1], [A2], [AK], [E], [Ex], [FL], [KMT], [KM1], [KM2], [K1], [KS], [LT], [M], [MP], [Pr], [Ra], [T1], [T2], [T4], [T5], and [V]. In this paper we give a simple and unified approach to the study of these polynomials. We also consider the case of the oscillator or Heisenberg group, leading to two more classes of polynomials.

In many cases such multivariable polynomials arose naturally from problems in physics or statistics. In view of the group theoretic interpretation, this is not surprising, since symmetry groups play an important role in these sciences. There are also close connections with spherical harmonics, as well as with the Wigner symbols (recoupling coefficients) occurring in quantum mechanics.

We will work with the analytic continuation of the holomorphic discrete series of $SU(1,1) \simeq SL(2,\mathbb{R})$. We realize these representations on the Hilbert spaces $\mathcal{A}^\nu$ ($\nu > 0$) of analytic functions on the complex unit disc, with the scalar product

$$\langle f, g\rangle = \sum_{k=0}^{\infty} \frac{k!}{(\nu)_k} \hat{f}(k)\overline{\hat{g}(k)},$$

where $f(z) = \sum \hat{f}(k) z^k$ and

$$(\nu)_k = \frac{\Gamma(\nu+k)}{\Gamma(\nu)} = \nu(\nu+1)\cdots(\nu+k-1)$$

*Received by the editors August 6, 1997; accepted for publication (in revised form) March 2, 1998; published electronically November 17, 1998.
http://www.siam.org/journals/sima/30-2/32568.html
†Department of Mathematics, University of Lund, Box 118, S-221 00 Lund, Sweden (hjalmar@maths.lth.se).

is the Pochhammer symbol. The norm is invariant under the transformations

$$f(z) \mapsto f\left(\frac{\alpha z+\beta}{\gamma z+\delta}\right)\frac{1}{(\gamma z+\delta)^\nu}, \qquad \begin{pmatrix}\alpha & \beta\\ \gamma & \delta\end{pmatrix} \in \mathrm{SU}(1,1).$$

This may be used to define a unitary representation of the universal covering group of $\mathrm{SU}(1,1)$ on each space $\mathcal{A}^\nu$, cf. [Sa] for the details.

We consider a Hilbert tensor product $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of such spaces. It decomposes under the group action as

$$\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n} \simeq \bigoplus_{s=0}^{\infty}\binom{n+s-2}{n-2}\mathcal{A}^{|\nu|+2s} \tag{1.1}$$

(where $|\nu|=\sum\nu_i$). A highest weight vector Q is an image of the constant function 1 under an intertwining embedding

$$\mathcal{A}^{|\nu|+2s} \to \mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}. \tag{1.2}$$

In our realization, this means that Q is a homogeneous polynomial of degree s, which may be expressed as a function of differences z_i-z_j of the coordinates. Equivalently, Q satisfies the identity

$$Q(az_1+b,\dots,az_n+b) = a^sQ(z_1,\dots,z_n), \qquad a,b\in\mathbb{C}. \tag{1.3}$$

In agreement with (1.1), the space of such polynomials has dimension $\binom{n+s-2}{n-2}$.

In [R], we introduced three transforms which we will denote here by T_1, T_2, and T_3. Their role is to send highest weight vectors to coupling coefficients. They are defined on $z^k\in\mathcal{A}^\nu$ by

$$T_1z^k(m) = (-1)^k\frac{(-m)_k}{(\nu)_k}, \quad T_2z^k(\xi) = \frac{\xi^k}{(\nu)_k}, \quad T_3z^k(X) = (-1)^k\frac{(\frac{\nu}{2}-iX)_k}{(\nu)_k} \tag{1.4}$$

and extended to power series by linearity. We will also denote by T_i the extension $T_i\otimes\cdots\otimes T_i$ to a tensor product $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$. Thus if we let

$$Q(z) = \sum_{|t|=s} c_t\, z_1^{t_1}\cdots z_n^{t_n}$$

be a highest weight vector in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ (throughout the paper we use multi-index notation freely), the three transforms of Q are given by

$$T_1Q(m_1,\dots,m_n) = (-1)^s\sum_{|t|=s} c_t\,\frac{(-m_1)_{t_1}\cdots(-m_n)_{t_n}}{(\nu_1)_{t_1}\cdots(\nu_n)_{t_n}}, \tag{1.5}$$

$$T_2Q(\xi_1,\dots,\xi_n) = \sum_{|t|=s} c_t\,\frac{\xi_1^{t_1}\cdots\xi_n^{t_n}}{(\nu_1)_{t_1}\cdots(\nu_n)_{t_n}}, \tag{1.6}$$

$$T_3Q(X_1,\dots,X_n) = (-1)^s\sum_{|t|=s} c_t\,\frac{(\frac{\nu_1}{2}-iX_1)_{t_1}\cdots(\frac{\nu_n}{2}-iX_n)_{t_n}}{(\nu_1)_{t_1}\cdots(\nu_n)_{t_n}}. \tag{1.7}$$

Note that the T_iQ are again polynomials of degree s, though only T_2Q is homogeneous. Moreover, T_1 and T_3 are connected by the equation

$$T_3Q(X) = T_1Q\left(iX-\frac{\nu}{2}\right). \tag{1.8}$$

One may also prove that

$$T_3Q(X) = (-1)^s\, T_3Q(-X), \tag{1.9}$$

which is not obvious from the definition.

In [R] we proved that if Q and Q' are two highest weight vectors, then the quantities

$$\sum_{m_1+\cdots+m_n=M} \frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}\, T_1Q(m_1,\ldots,m_n)\,\overline{T_1Q'(m_1,\ldots,m_n)}$$

for $M = 0, 1, 2, \ldots$,

$$\int_{\xi\in\mathbb{R}^n_+\,:\,|\xi|=C} T_2Q(\xi)\,\overline{T_2Q'(\xi)}\,\xi_1^{\nu_1-1}\cdots\xi_n^{\nu_n-1}\,d\xi,$$

where $|\xi| = \xi_1 + \cdots + \xi_n$ and $C > 0$ is arbitrary, and

$$\int_{X\in\mathbb{R}^n\,:\,|X|=C} T_3Q(X)\,\overline{T_3Q'(X)}\,\left|\Gamma\left(\frac{\nu_1}{2}+iX_1\right)\right|^2\cdots\left|\Gamma\left(\frac{\nu_n}{2}+iX_n\right)\right|^2 dX,$$

where $C \in \mathbb{R}$ is arbitrary, are all proportional to the scalar product

$$\langle Q, Q'\rangle_{\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}}.$$

In particular, given an orthogonal basis for the highest weight vectors, our transforms give us three systems of orthogonal polynomials. The form of the orthogonality relations suggests that one eliminate one variable and consider, for instance, for fixed M,

$$T_1Q(m_1,\ldots,m_{n-1}, M-m_1-\cdots-m_{n-1})$$

as a polynomial in $n-1$ variables. This is how these polynomials usually occur in the literature.

In the case when all the ν_i are half-integers, there is a connection with spherical harmonics. If we introduce the polynomial

$$R(x_1,\ldots,x_{2|\nu|}) = (T_2Q)(x_1^2+\cdots+x_{2\nu_1}^2,\ldots,x_{2(\nu_1+\cdots+\nu_{n-1})+1}^2+\cdots+x_{2|\nu|}^2),$$

then the orthogonality relations for T_2Q give corresponding orthogonality relations for R on the unit sphere of $\mathbb{R}^{2\nu}$. Moreover, the conditions on Q ensure that R is harmonic. In fact one obtains precisely the space of harmonic polynomials of degree s invariant with respect to the subgroup

$$\mathrm{O}(2\nu_1)\times\cdots\times\mathrm{O}(2\nu_n)\subseteq\mathrm{O}(2|\nu|),$$

called *polyspherical harmonics* [V]. In [R] we gave an explanation of this, which involved the "Howe dual pair" $\mathrm{Mp}(1)\times\mathrm{O}(k)\subseteq\mathrm{Mp}(k)$.

The plan of the paper is as follows. In section 2 we describe some special cases of our polynomials, in particular those that we have found in the literature. In section 3 we review some fundamental facts on matrix elements of our representation due to Basu and Wolf [BW1]. In section 4 we obtain the polynomials T_iQ as coupling coefficients for our representation. The proof is based on certain factorizations of the Fourier transforms on $\mathcal{A}^\nu$ with respect to one-parameter subgroups. This is

different from the approach in [R]. In section 5 we prove convolution formulas for our polynomials. These generalize formulas recently found by Koelink and Van der Jeugt [KV1]. In section 6 we introduce some polynomials in $2n$ variables which we call coupling kernels. Whereas the coupling coefficients T_iQ are connected with matrix elements of the intertwining embeddings (1.2) (or of the corresponding projections $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}\to\mathcal{A}^{|\nu|+2s}$), the coupling kernels are connected with matrix elements for the projection of $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ onto the isotypic subspaces equivalent to $\binom{n+s-2}{n-2}\mathcal{A}^{|\nu|+2s}$ in (1.1). In section 7 we study formal properties of the coupling kernels. These include some linearization formulas, which generalize the classical Burchnall–Chaundy formula. In section 8 we prove that the coupling kernels may be expressed in terms of certain generalized hypergeometric functions. Finally, in section 9 we give the analogues of our results when $\mathrm{SU}(1,1)$ is replaced by the Heisenberg group and the spaces $\mathcal{A}^\nu$ are replaced by Fock spaces of entire functions.

2. Special cases. We will now describe some special cases of our polynomials. We have found many occurrences of them in the literature after writing [R], so here we will try to give more complete references. We first recall the definition of the generalized hypergeometric function

$$ {}_pF_q\left(\left.\begin{matrix}a_1,\dots,a_p\\ b_1,\dots,b_q\end{matrix}\right|x\right)=\sum_{k=0}^{\infty}\frac{(a_1)_k\cdots(a_p)_k\,x^k}{(b_1)_k\cdots(b_q)_k\,k!}. $$

When a and k are vectors, we write

$$ (a)_k=(a_1)_{k_1}\cdots(a_n)_{k_n},\qquad k!=k_1!\cdots k_n!. $$

We will also encounter Karlsson's generalized Kampé de Fériet functions [Ka], [SK], which we may then write, with a slight variation of Karlsson's notation, as

$$ \begin{aligned} &F^{p:q}_{r:s}\left(\left.\begin{matrix}a_1,\dots,a_p\ :\ b_1,\dots,b_q\\ c_1,\dots,c_r\ :\ d_1,\dots,d_s\end{matrix}\right|x\right)\\ &=\sum_{k_1,\dots,k_n=0}^{\infty}\frac{(a_1)_{|k|}\cdots(a_p)_{|k|}(b_1)_k\cdots(b_q)_k\,x^k}{(c_1)_{|k|}\cdots(c_r)_{|k|}(d_1)_k\cdots(d_s)_k\,k!}, \end{aligned} \tag{2.1} $$

where the a_i and c_i are scalar parameters while the b_i, d_i, and x are vectors of the same dimension n. For $n=1$, $F^{p:q}_{r:s}$ reduces to the function ${}_{p+q}F_{r+s}$.

2.1. The bilinear case. The simplest case is when $n=2$. Then the space of highest weight vectors of each degree s is one-dimensional and spanned by the single element

$$ Q(z_1,z_2)=(z_1-z_2)^s. $$

In this case, the polynomials T_1Q are Clebsch–Gordan coefficients, which may be expressed in terms of Hahn polynomials. In fact,

$$ \begin{aligned} T_1Q(m_1,m_2)&=\sum_{k=0}^{s}\binom{s}{k}(-1)^k\frac{(-m_1)_k(-m_2)_{s-k}}{(\nu_1)_k(\nu_2)_{s-k}}\\ &=\frac{(-m_1-m_2)_s}{(\nu_2)_s}\,Q_s(m_1;\nu_1-1,\nu_2-1;m_1+m_2), \end{aligned} $$

where Q_s is the Hahn polynomial of degree s in the notation of [VK]. The polynomials T_2Q and T_3Q are Clebsch–Gordan coefficients with respect to continuous bases; they may be expressed in terms of Jacobi polynomials and "Hahn polynomials of an imaginary argument," respectively; cf. [MR], [BW2], [P2], and [VK, section 8.7]. Still in the notation of [VK], we have

$$\begin{aligned} T_2Q(\xi_1,\xi_2) &= \sum_{k=0}^{s} \binom{s}{k} (-1)^{s-k} \frac{\xi_1^k\, \xi_2^{s-k}}{(\nu_1)_k (\nu_2)_{s-k}} \\ &= \frac{(-1)^s s!}{(\nu_1)_s (\nu_2)_s} (\xi_1+\xi_2)^s P_s^{(\nu_1-1,\nu_2-1)}\left(\frac{\xi_2-\xi_1}{\xi_1+\xi_2}\right), \\ T_3Q(X_1,X_2) &= \sum_{k=0}^{s} \binom{s}{k} (-1)^k \frac{(\frac{\nu_1}{2}-iX_1)_k (\frac{\nu_2}{2}-iX_2)_{s-k}}{(\nu_1)_k (\nu_2)_{s-k}} \\ &= \frac{(\frac{\nu_1+\nu_2}{2}-i(X_1+X_2))_s}{(\nu_2)_s}\, q_s(X_1; \tfrac{\nu_1}{2}, \tfrac{\nu_2}{2} - i(X_1+X_2)). \end{aligned}$$

Even in this case, the present paper gives new and simple deductions of properties of these polynomials.

2.2. Polynomials of Appell-type. In the case of general n, the polynomials

$$Q_t(z) = (z_1-z_n)^{t_1}\cdots(z_{n-1}-z_n)^{t_{n-1}}, \qquad t_1+\cdots+t_{n-1}=s, \tag{2.2}$$

form a basis for the space of highest weight vectors of degree s. Expanding Q_t by the binomial theorem, one finds that the polynomials T_iQ_t may be expressed in terms of Karlsson's generalized Kampé de Fériet functions (2.1) with the vector parameters of dimension $n-1$. In fact, one has

$$\begin{aligned} &T_1Q_t(m_1,\dots,m_n) \\ &= \frac{(-m_n)_s}{(\nu_n)_s} \sum_{k_1,\dots,k_{n-1}=0}^{\infty} \frac{(1-\nu_n-s)_{|k|}(-t_1)_{k_1}\cdots(-t_{n-1})_{k_{n-1}}(-m_1)_{k_1}\cdots(-m_{n-1})_{k_{n-1}}}{(1+m_n-s)_{|k|}(\nu_1)_{k_1}\cdots(\nu_{n-1})_{k_{n-1}} k_1!\cdots k_{n-1}!} \\ &= \frac{(-m_n)_s}{(\nu_n)_s} F^{1:2}_{1:1}\left(\begin{matrix} 1-\nu_n-s & : & -t, -(m_1,\dots,m_{n-1}) \\ 1+m_n-s & : & (\nu_1,\dots,\nu_{n-1}) \end{matrix}\,\middle|\, 1\right) \end{aligned}$$

(where, as is customary, we write the variable of the $F^{1:2}_{1:1}$-function as $1=(1,\dots,1)$). By (1.8), a similar expression is valid for T_3Q_t, while T_2Q_t is given by

$$\begin{aligned} &T_2Q_t(\xi_1,\dots,\xi_n) \\ &= \frac{(-\xi_n)^s}{(\nu_n)_s} F^{1:1}_{0:1}\left(\begin{matrix} 1-\nu_n-s & : & -t \\ - & : & (\nu_1,\dots,\nu_{n-1}) \end{matrix}\,\middle|\, -\frac{\xi_1}{\xi_n},\dots,-\frac{\xi_{n-1}}{\xi_n}\right) \end{aligned} \tag{2.3}$$

(Karlsson's $F^{1:1}_{0:1}$-function is also known as the Lauricella F_A-function).

2.3. Polynomials biorthogonal to the polynomials of Appell-type. The family (2.2) is obviously nonorthogonal for $n \geq 3$. Thus it is a problem to find the dual basis. Applying the T_i to these two families then gives biorthogonal polynomial systems. Though we have found no very simple expression for the basis dual to (2.2), the polynomials obtained after applying the T_i are also expressible in terms of Karlsson's functions; cf. Corollary 6.2 below.

The polynomials (2.3) were introduced by Appell [A1] for $n = 3$. The dual basis was found by him [A2] when $\nu_3 = 1$ and by Engelis [E] and, independently, by Fackerell and Littler [FL] in general. (For the work of Appell, see also the book [AK].) Lam and Tratnik [LT] generalized these polynomials to general n. The corresponding spherical harmonics occur in [KMT]. In the special case when all the ν_i equal 1, these harmonics go back to Hermite, Didon, and Kampé de Fériet; cf. again [AK]. The biorthogonal systems obtained using T_1 were introduced by Rahman [Ra] for $n = 3$ and by Tratnik [T2] in general. For the transform T_3, the corresponding polynomials are also due to Tratnik [T1].

2.4. Polynomials of Proriol-type. There seems to be no canonical choice of an orthogonal basis in the space of highest weight vectors. There is, however, a general procedure, known as binary coupling (the coupling of angular momenta in quantum physics), that allows one to construct many such bases. The corresponding polynomials T_iQ are given by products of $n-1$ factors, each factor being a Clebsch–Gordan coefficient, that is a Hahn, Jacobi, or continuous Hahn polynomial.

As an example, let us consider the case $n = 3$. Let $K_s^{\nu_1,\nu_2}$ denote the intertwining embedding

$$\mathcal{A}^{\nu_1+\nu_2+2s} \to \mathcal{A}^{\nu_1} \otimes \mathcal{A}^{\nu_2}$$

such that $K_s^{\nu_1,\nu_2}1 = (z_1 - z_2)^s$. Then the polynomials

$$Q_{st} = (K_s^{\nu_1,\nu_2} \otimes \mathrm{id}) \circ K_t^{\nu_1+\nu_2+2s,\nu_3}1, \qquad s + t = N$$

form an orthogonal basis for the space of highest weight vectors in $\mathcal{A}^{\nu_1} \otimes \mathcal{A}^{\nu_2} \otimes \mathcal{A}^{\nu_3}$ of degree N. In [R] we found that

$$\begin{aligned}
Q_{st}(z) &= (z_1 - z_2)^s \sum_{i+j=t} \frac{t!}{i!\,j!} \frac{(\nu_1 + s)_i (\nu_2 + s)_j}{(\nu_1 + \nu_2 + 2s)_t} (z_1 - z_3)^i (z_2 - z_3)^j, \\
T_1Q_{st}(m) &= \frac{(-m_1 - m_2)_s (s - m_1 - m_2 - m_3)_t}{(\nu_2)_s (\nu_3)_t} Q_s(m_1; \nu_1 - 1, \nu_2 - 1; m_1 + m_2) \\
&\quad \times Q_t(m_1 + m_2 - s; \nu_1 + \nu_2 + 2s - 1, \nu_3 - 1; m_1 + m_2 + m_3 - s), \\
T_2Q_{st}(\xi) &= \frac{(-1)^{s+t} s!\, t! (\xi_1 + \xi_2)^s (\xi_1 + \xi_2 + \xi_3)^t}{(\nu_1)_s (\nu_2)_s (\nu_1 + \nu_2 + 2s)_t (\nu_3)_t} \\
&\quad \times P_s^{(\nu_1 - 1, \nu_2 - 1)} \left(\frac{\xi_2 - \xi_1}{\xi_1 + \xi_2} \right) P_t^{(\nu_1 + \nu_2 + 2s - 1, \nu_3 - 1)} \left(\frac{\xi_3 - \xi_1 - \xi_2}{\xi_1 + \xi_2 + \xi_3} \right), \\
T_3Q_{st}(X) &= \frac{(\frac{\nu_1 + \nu_2}{2} - i(X_1 + X_2))_s (\frac{\nu_1 + \nu_2 + \nu_3 + 2s}{2} - i(X_1 + X_2 + X_3))_t}{(\nu_2)_s (\nu_3)_t} \\
&\quad \times q_s \left(X_1; \frac{\nu_1}{2}, \frac{\nu_2}{2} - i(X_1 + X_2) \right) \\
&\quad \times q_t \left(X_1 + X_2; \frac{\nu_1 + \nu_2 + 2s}{2}, \frac{\nu_3}{2} - i(X_1 + X_2 + X_3) \right).
\end{aligned}$$

The polynomials T_2Q_{st}, with Q_{st} as above, were introduced by Proriol in [Pr]. They were independently obtained and applied to genetics by Karlin and McGregor in [KM1] and have also found applications in physics [M], [MP]. See the survey [K1] for these and other two-variable analogues of Jacobi polynomials. The spherical harmonics constructible by this method were, for general n, introduced in [V], cf.

also [VK, section 10.5], and generalized further in [KMT]. For the special coupling procedure

$$Q_{s_1,\dots,s_{n-1}} = (K_{s_1}^{\nu_1,\nu_2} \otimes \mathrm{id}) \circ (K_{s_2}^{\nu_1+\nu_2+2s_1,\nu_3} \otimes \mathrm{id}) \circ \cdots \circ K_{s_{n-1}}^{\nu_1+\cdots+\nu_{n-1}+2(s_1+\cdots+s_{n-2}),\nu_n} 1,$$

the polynomials $T_1 Q_{s_1,\dots,s_{n-1}}$ were defined by Karlin and McGregor in [KM2], cf. also [T5]; again they appear in the context of genetics. The corresponding polynomials $T_3 Q_{s_1,\dots,s_{n-1}}$ were introduced by Tratnik in [T4]. An interesting product formula for the polynomials $T_2 Q_{s_1,\dots,s_{n-1}}$ is given in [KS].

3. Matrix elements. In this section we write down some basic facts, due to Basu and Wolf [BW1], cf. also [K2], about matrix elements of our representation. Whereas these authors work with the oscillator realization on $\mathrm{L}^2(\mathbb{R}_+)$, we will reformulate their results in terms of the analytic function spaces $\mathcal{A}^\nu$. This has computational advantages, allowing us to work with power series expansions.

Let $\mathbf{D}$ be the unit disc of $\mathbb{C}$ and $G \simeq \mathrm{SU}(1,1)/\{\pm 1\}$ be the group of conformal automorphisms of $\mathbf{D}$. There are three conjugacy classes of one-parameter subgroups in G. They may be described in terms of the fixed point sets in $\mathbb{C} \cup \{\infty\}$ of the corresponding Möbius maps. There are the *elliptic* elements, which fix one point in $\mathbf{D}$ and one in the outside, the *hyperbolic* elements, which fix two points on the boundary of $\mathbf{D}$, and finally the *parabolic* elements, which have a single fixed point on the boundary.

We choose a representative subgroup in each conjugacy class, namely, the elliptic subgroup

$$\phi_t^1 = \begin{pmatrix} e^{it/2} & 0 \\ 0 & e^{-it/2} \end{pmatrix},$$

corresponding to rotations $z \mapsto e^{it} z$ of the disc, the parabolic subgroup

$$\phi_t^2 = \begin{pmatrix} 1 + it/2 & it/2 \\ -it/2 & 1 - it/2 \end{pmatrix},$$

and the hyperbolic subgroup

$$\phi_t^3 = \begin{pmatrix} \cosh(t/2) & -\sinh(t/2) \\ -\sinh(t/2) & \cosh(t/2) \end{pmatrix}.$$

One may check that

$$\phi_s^i \phi_t^i = \phi_{s+t}^i, \qquad i = 1, 2, 3.$$

Note that if we choose ψ as the Cayley transform

$$\psi(z) = \frac{i - z}{i + z}$$

of the upper half-plane onto $\mathbf{D}$, then

$$(\psi^{-1} \phi_t^2 \psi)(z) = z + t,$$

so we may realize the parabolic subgroup as translations of the half-plane. Similarly

$$(\gamma^{-1} \phi_t^3 \gamma)(z) = z + t,$$

where

$$\gamma(z)=\frac{i-e^z}{i+e^z},$$

which realizes the hyperbolic subgroup as translations of the strip $\{z\,|\,0<\mathrm{Im}z<\pi\}$. (Hyperbolic transformations may also be realized as dilatations of the half-plane.)

To clarify the group theoretic meaning of our formulas, we employ certain notational conventions, which we will now state explicitly. Various objects connected with the subgroup ϕ_t^i $(i=1,2,3)$ will be labelled with the index i. The letter z will denote variables ranging in the disc (or polydisc). We will be concerned with three versions of the Fourier transform, one for each subgroup. For the elliptic subgroup, the variable dual to z is discrete; it will be denoted by k,l,m. For the parabolic and hyperbolic subgroup, the dual variables are continuous; they will be denoted by ξ,η and X,Y, respectively. The equations (1.5), (1.6), and (1.7) are examples of these conventions.

Recall now the function spaces $\mathcal{A}^\nu$ $(\nu>0)$ with scalar product

$$\langle f,g\rangle=\sum_{k=0}^{\infty}\frac{k!}{(\nu)_k}\hat{f}(k)\overline{\hat{g}(k)}.$$

We will use this notation also when f or g is not in $\mathcal{A}^\nu$, provided that the sum on the right-hand side converges. This space has the reproducing kernel

$$k_w(z)=\sum_{k=0}^{\infty}\frac{(\nu)_k}{k!}(z\bar{w})^k=\frac{1}{(1-z\bar{w})^\nu},$$

that is,

$$f(w)=\langle f,k_w\rangle,\qquad f\in\mathcal{A}^\nu,\ w\in\mathbb{D}.$$

We denote the group action by

$$(f*\phi)(z)=f(\phi(z))\phi'(z)^{\frac{\nu}{2}}=f\left(\frac{\alpha z+\beta}{\bar{\beta}z+\bar{\alpha}}\right)\frac{1}{(\bar{\beta}z+\bar{\alpha})^\nu},$$

where $\phi\in G$ is given by

$$\phi(z)=\frac{\alpha z+\beta}{\bar{\beta}z+\bar{\alpha}},\qquad |\alpha|^2-|\beta|^2=1.$$

We also write

$$\mathcal{U}_\phi(f)=f*\phi^{-1}$$

for the corresponding left action. Because of the indeterminacy of the power function, one must pass to the universal covering group to get a unitary representation on each space $\mathcal{A}^\nu$. (If $\nu\in 2\mathbb{Z}/k$ for an integer k, the k-fold cover is sufficient.)

It is easy to find joint eigenfunctions of our subgroups. Namely, since $e^{i\xi z}$ are eigenfunctions of the translations, it follows that

$$e^{i\xi\psi^{-1}(z)}((\psi^{-1})'(z))^{\frac{\nu}{2}},$$

with ψ as above, are eigenfunctions of ϕ_t^2. It will be convenient to normalize them so that they take the value 1 at the origin. This gives the functions

$$e_\xi^2(z) = \frac{1}{(1+z)^\nu} \exp\left(\frac{\xi z}{1+z}\right),$$

which indeed satisfy

$$e_\xi^2 * \phi_t^2 = e^{it\xi} e_\xi^2.$$

Replacing ψ by γ in this argument leads to the functions

$$e_X^3(z) = \frac{2^{\frac{\nu}{2}}}{(1-z)^{\frac{\nu}{2}-iX}(1+z)^{\frac{\nu}{2}+iX}}$$

(the factor $2^{\frac{\nu}{2}}$ being chosen for convenience), which satisfy

$$e_X^3 * \phi_t^3 = e^{itX} e_X^3.$$

Finally we write

$$e_k^1(z) = z^k$$

for the eigenfunctions of the elliptic subgroup. We call $(e_k^1)_{k=0}^\infty$ the elliptic basis of $\mathcal{A}^\nu$, while $(e_\xi^2)_{\xi>0}$ and $(e_X^3)_{X\in\mathbb{R}}$ are called the parabolic and the hyperbolic (generalized) basis, respectively.

Note that, whereas clearly e_k^1 is an element of $\mathcal{A}^\nu$, this is not true for e_ξ^2 or e_X^3. However, one may still define $\langle f, e_x^i\rangle$ $(i = 2, 3)$ for f in a dense subspace of $\mathcal{A}^\nu$. Natural choices are the space of polynomials and the linear hull of all reproducing kernels k_z, $z \in \mathbf{D}$. With either of these choices the following Plancherel theorem is true.

PROPOSITION 3.1. *The operator* $(\mathcal{F}_2 f)(\xi) = \langle f, e_\xi^2\rangle$ *extends to a Hilbert space isomorphism*

$$\mathcal{F}_2 : \mathcal{A}^\nu \to \mathrm{L}^2\left(\mathbb{R}_+, \frac{1}{\Gamma(\nu)}\, e^{-\xi}\xi^{\nu-1}\, d\xi\right).$$

Similarly, the operator $(\mathcal{F}_3 f)(X) = \langle f, e_X^3\rangle$ *extends to an isomorphism*

$$\mathcal{F}_3 : \mathcal{A}^\nu \to \mathrm{L}^2\left(\mathbb{R}, \frac{1}{2\pi\Gamma(\nu)}\left|\Gamma\left(\frac{\nu}{2}+iX\right)\right|^2 dX\right).$$

The transforms $\mathcal{F}_2$ and $\mathcal{F}_3$ are connected to the classical Fourier transform. For instance $\mathcal{F}_2 f$ is essentially the Fourier transform of the boundary value distribution of the function $f(\psi(z))\psi'(z)^{\nu/2}$ defined in the upper half-plane (they differ by a multiplicative factor). The proposition may be proved using this observation, cf. [P2], or alternatively using the orthogonality relations for Laguerre and Meixner–Pollaczek polynomials (whose role is made clear below).

By polarization, we have

$$\begin{aligned}\langle f, g\rangle &= \sum_{k=0}^\infty \frac{(\nu)_k}{k!}\langle f, e_k^1\rangle\langle e_k^1, g\rangle \\ &= \frac{1}{\Gamma(\nu)}\int_0^\infty \langle f, e_\xi^2\rangle\langle e_\xi^2, g\rangle\, e^{-\xi}\xi^{\nu-1}\, d\xi \\ &= \frac{1}{2\pi\Gamma(\nu)}\int_{-\infty}^\infty \langle f, e_X^3\rangle\langle e_X^3, g\rangle \left|\Gamma\left(\frac{\nu}{2}+iX\right)\right|^2 dX\end{aligned} \tag{3.1}$$

with a suitable interpretation of the last two expressions. Choosing $g = k_z$, it follows that

$$\begin{aligned} f(z) &= \sum_{k=0}^{\infty} \frac{(\nu)_k}{k!} (\mathcal{F}_1 f)(k) e^1_k(z) \\ &= \frac{1}{\Gamma(\nu)} \int_0^{\infty} (\mathcal{F}_2 f)(\xi) e^2_\xi(z) e^{-\xi} \xi^{\nu-1}\, d\xi \\ &= \frac{1}{2\pi\Gamma(\nu)} \int_{-\infty}^{\infty} (\mathcal{F}_3 f)(X) e^3_X(z) \left|\Gamma\left(\frac{\nu}{2} + iX\right)\right|^2 d\xi, \end{aligned} \tag{3.2}$$

where

$$\mathcal{F}_1 f(k) = \langle f, e^1_k \rangle = \frac{k!}{(\nu)_k} \hat{f}(k).$$

We now introduce the *matrix elements*

$$K^{ij}(x, y, \phi) = \langle e^i_x, e^j_y * \phi \rangle, \qquad i, j = 1, 2, 3,$$

where $\phi \in G$ is such that this makes sense. By unitarity,

$$K^{ij}(x, y, \phi) = \overline{K^{ji}(y, x, \phi^{-1})},$$

so there are six cases to consider. As remarked above, the K^{ij} were computed by Basu and Wolf [BW1]. We summarize their results in the following proposition.

PROPOSITION 3.2. *The matrix elements of $\mathcal{A}^\nu$ are given by*

$$K^{11}(k, l, \phi) = \begin{cases} \dfrac{\beta^k \bar{\beta}^l}{\alpha^{\nu+k+l}} (-1)^k \, {}_2F_1\left(\begin{matrix} -k, -l \\ \nu \end{matrix} \,\middle|\, -\dfrac{1}{|\beta|^2} \right) & \beta \neq 0, \\ \dfrac{k!}{(\nu)_k \alpha^{\nu+2k}}\, \delta_{kl} & \beta = 0, \end{cases}$$

$$K^{12}(k, \xi, \phi) = \frac{(-1)^k (\bar{\alpha} + \beta)^k}{(\alpha + \bar{\beta})^{\nu+k}} \exp\left(\frac{\bar{\beta}\xi}{\bar{\beta} + \alpha}\right) {}_1F_1\left(\begin{matrix} -k \\ \nu \end{matrix} \,\middle|\, \frac{\xi}{|\alpha + \bar{\beta}|^2} \right),$$

$$K^{13}(k, X, \phi) = \frac{2^{\frac{\nu}{2}} (\bar{\alpha} - \beta)^k}{(\alpha + \bar{\beta})^{\frac{\nu}{2} - iX} (\alpha - \bar{\beta})^{\frac{\nu}{2} + iX + k}} \, {}_2F_1\left(\begin{matrix} -k, \frac{\nu}{2} - iX \\ \nu \end{matrix} \,\middle|\, \frac{2}{(\alpha + \bar{\beta})(\bar{\alpha} - \beta)} \right),$$

$$\begin{aligned} K^{22}(\xi, \eta, \phi) = {} & \frac{1}{(2i\mathrm{Im}(\alpha - \beta))^\nu} \exp\left(-\frac{(\bar{\alpha} + \beta)\xi + (\bar{\alpha} - \bar{\beta})\eta}{2i\mathrm{Im}(\alpha - \beta)} \right) \\ & \times {}_0F_1\left(\begin{matrix} - \\ \nu \end{matrix} \,\middle|\, -\frac{\xi\eta}{4(\mathrm{Im}(\alpha - \beta))^2} \right), \end{aligned}$$

$$\begin{aligned} K^{23}(\xi, X, \phi) = {} & \frac{2^{\frac{\nu}{2}}}{(2\mathrm{Re}(\alpha - \beta))^{\frac{\nu}{2} + iX} (2i\mathrm{Im}(\alpha - \beta))^{\frac{\nu}{2} - iX}} \exp\left(\frac{(\bar{\alpha} - \beta)\xi}{2\mathrm{Re}(\alpha - \beta)} \right) \\ & \times {}_1F_1\left(\begin{matrix} \frac{\nu}{2} - iX \\ \nu \end{matrix} \,\middle|\, \frac{i\xi}{2\mathrm{Re}(\alpha - \beta)\mathrm{Im}(\alpha - \beta)} \right), \end{aligned}$$

$$\begin{aligned} K^{33}(X, Y, \phi) = {} & \frac{(i\mathrm{Im}(\alpha + \beta))^{i(X-Y)}}{(\mathrm{Re}(\alpha - \beta))^{\frac{\nu}{2} + iX} (\mathrm{Re}(\alpha + \beta))^{\frac{\nu}{2} - iY}} \\ & \times {}_2F_1\left(\begin{matrix} \frac{\nu}{2} + iX, \frac{\nu}{2} - iY \\ \nu \end{matrix} \,\middle|\, \frac{1}{\mathrm{Re}(\alpha + \beta)\mathrm{Re}(\alpha - \beta)} \right), \end{aligned}$$

where $\phi(z) = (\alpha z + \beta)/(\bar{\beta} z + \bar{\alpha})$, $|\alpha|^2 - |\beta|^2 = 1$.

One may check that these are well defined up to a choice of $(\phi')^{\nu/2}$. The matrix elements may be identified with classical orthogonal functions as follows, cf. [K2]:

K^{11}	:	Meixner polynomials,
K^{12}	:	Laguerre polynomials,
K^{13}	:	Meixner–Pollaczek polynomials,
K^{22}	:	Bessel functions,
K^{23}	:	Laguerre functions,
K^{33}	:	Meixner–Pollaczek functions.

By inserting basis elements $e_x^i * \phi$ into (3.1), one may deduce a host of addition theorems and orthogonality relations for these functions, cf. [VK, section 7.7].

The proposition does not give the matrix elements K^{22}, K^{23}, or K^{33} for $\phi = \mathrm{id}$. However, by (3.2) it is natural to define

$$K^{22}(\xi, \eta, \mathrm{id}) = \Gamma(\nu) e^{\xi} \xi^{1-\nu} \delta(\xi - \eta), \tag{3.3}$$

$$K^{33}(X, Y, \mathrm{id}) = 2\pi\Gamma(\nu) \left|\Gamma\left(\frac{\nu}{2} + iX\right)\right|^{-2} \delta(X - Y), \tag{3.4}$$

where δ is the Dirac measure. Similarly one may define

$$K^{23}(\xi, X, \mathrm{id}) = \frac{\Gamma(\nu) e^{\frac{\xi}{2}}}{\Gamma(\frac{\nu}{2} + iX) 2^{iX}} \, \xi^{-\frac{\nu}{2} + iX}$$

since this kernel has the property

$$e_X^3(z) = \frac{1}{\Gamma(\nu)} \int_0^\infty \overline{K^{23}(\xi, X, \mathrm{id})} \, e_\xi^2(z) e^{-\xi} \xi^{\nu-1} \, d\xi.$$

(This reduces to the usual definition of the Γ-function after a change of variables.) These identities can be made precise by extending the transforms $\mathcal{F}_i$ to suitable classes of distributions.

As for the proof, we remark that in the present realization, *all* of Proposition 3.2 may be deduced from Meixner's expansion formula [Me]

$$\begin{aligned} &\sum_{k=0}^{\infty} \frac{(\nu)_k}{k!} \, {}_2F_1\left(\begin{matrix} -k, \alpha \\ \nu \end{matrix} \,\middle|\, x \right) {}_2F_1\left(\begin{matrix} -k, \beta \\ \nu \end{matrix} \,\middle|\, y \right) z^k \\ &= \frac{1}{(1-z)^{\nu-\alpha-\beta}(1-z+xz)^{\alpha}(1-z+yz)^{\beta}} \, {}_2F_1\left(\begin{matrix} \alpha, \beta \\ \nu \end{matrix} \,\middle|\, \frac{xyz}{(1-z+xz)(1-z+yz)} \right). \end{aligned} \tag{3.5}$$

The case $x = 0$ gives the expressions for K^{11} and K^{13}. To proceed, one may use the expansion

$$K^{33}(X, Y, \phi) = \sum_{k=0}^{\infty} \frac{(\nu)_k}{k!} \, \overline{K^{13}(k, X, \mathrm{id})} \, K^{13}(k, Y, \phi),$$

which follows from the first identity in (3.1). The expression for K^{33} given in Proposition 3.2 now follows from the general case of Meixner's formula.

The matrix elements involving the parabolic subgroup are computed similarly, using two degenerate cases of Meixner's formula. If we replace y by y/β in (3.5) and let β tend to infinity, we obtain the expansion

$$\begin{aligned}&\sum_{k=0}^{\infty}\frac{(\nu)_k}{k!}\,{}_2F_1\left(\begin{matrix}-k,\alpha\\ \nu\end{matrix}\,\middle|\,x\right){}_1F_1\left(\begin{matrix}-k\\ \nu\end{matrix}\,\middle|\,y\right)z^k\\ &=\frac{1}{(1-z)^{\nu-\alpha}(1-z+xz)^{\alpha}}\exp\left(-\frac{yz}{1-z}\right){}_1F_1\left(\begin{matrix}\alpha\\ \nu\end{matrix}\,\middle|\,\frac{xyz}{(1-z+xz)(1-z)}\right),\end{aligned}\tag{3.6}$$

due to Weisner. Repeating this once more gives the Hardy–Hille formula

$$\begin{aligned}&\sum_{k=0}^{\infty}\frac{(\nu)_k}{k!}\,{}_1F_1\left(\begin{matrix}-k\\ \nu\end{matrix}\,\middle|\,x\right){}_1F_1\left(\begin{matrix}-k\\ \nu\end{matrix}\,\middle|\,y\right)z^k\\ &=\frac{1}{(1-z)^{\nu}}\exp\left(-\frac{z(x+y)}{1-z}\right){}_0F_1\left(\begin{matrix}-\\ \nu\end{matrix}\,\middle|\,\frac{xyz}{(1-z)^2}\right).\end{aligned}$$

We refer to [SM] for a large number of proofs of these formulas.

4. Coupling coefficients. In this section we will obtain the polynomials T_iQ as coupling coefficients of our representation. We consider transforms of the type

$$\mathcal{F}_i^{\phi}f(x)=\mathcal{F}_i\,\mathcal{U}_{\phi}f(x)=\langle f,e_x^i*\phi\rangle,\qquad i=1,2,3,\ \phi\in G,$$

that is, the Fourier transform with respect to an arbitrary one-parameter subgroup of G. We will factorize these into operators which interact nicely with the Lie algebra action. More precisely, we will find factorizations of the type

$$\mathcal{F}_i^{\phi}=M_i\,T_i\,\delta_a\tau_b,$$

where δ_a and τ_b are the dilatation and translation operators defined by

$$\delta_af(z)=f(az),\qquad \tau_bf(z)=f(z+b),$$

M_i are the multiplication operators

$$M_if(x)=\overline{e_x^i*\phi(0)}f(x)=(\mathcal{F}_i^{\phi}1)(x)f(x),$$

and T_i are the transforms defined in (1.4).

We will use the explicit expressions for the matrix elements given in Proposition 3.2 to obtain our factorizations. Note, however, that we only need them in the easiest special case, corresponding to $x=0$ in (3.5) and (3.6).

PROPOSITION 4.1. *In the notation above, we have for* $\phi=\left(\begin{smallmatrix}\alpha&\beta\\ \bar\beta&\bar\alpha\end{smallmatrix}\right)$ *the following factorizations of densely defined operators*:

$$\begin{aligned}\mathcal{F}_1^{\phi}&=M_1\,T_1\,\delta_{1/\alpha\bar\beta}\,\tau_{-\beta/\alpha}\qquad(\beta\neq0),\\ \mathcal{F}_2^{\phi}&=M_2\,T_2\,\delta_{1/(\alpha+\bar\beta)^2}\,\tau_{-(\bar\alpha+\beta)/(\alpha+\bar\beta)},\\ \mathcal{F}_3^{\phi}&=M_3\,T_3\,\delta_{2/(\alpha+\bar\beta)(\alpha-\bar\beta)}\,\tau_{(\bar\alpha-\beta)/(\alpha-\bar\beta)}.\end{aligned}\tag{4.1}$$

Proof. Test the action of both sides on the function $e_k^1(z)=z^k$. In general,

$$(\delta_a\tau_b\,e_k^1)(z)=(az+b)^k=b^k\sum_{j=0}^{\infty}\frac{(-k)_j}{j!}\left(-\frac{az}{b}\right)^j.$$

Thus one has for instance

$$\begin{aligned} M_1\,T_1\,\delta_{1/\alpha\bar\beta}\,\tau_{-\beta/\alpha}\,e_k^1(m) &= \frac{\bar\beta^m}{\alpha^{\nu+m}}\,T_1\left(\left(-\frac{\beta}{\alpha}\right)^k\sum_{j=0}^{\infty}\frac{(-k)_j}{j!}\left(\frac{z}{|\beta|^2}\right)^j\right)(m)\\ &= \frac{\bar\beta^m(-\beta)^k}{\alpha^{\nu+m+k}}\sum_{j=0}^{\infty}\frac{(-k)_j(-m)_j}{j!(\nu)_j}\left(-\frac{1}{|\beta|^2}\right)^j\\ &= \frac{\bar\beta^m(-\beta)^k}{\alpha^{\nu+k+m}}\,{}_2F_1\left(\begin{matrix}-k,-m\\ \nu\end{matrix}\,\middle|\,-\frac{1}{|\beta|^2}\right)\\ &= K^{11}(k,m,\phi) = \mathcal{F}_1^\phi e_k^1(m), \end{aligned}$$

which proves (4.1). The other two identities follow similarly. □

Now let us apply these factorizations to a general tensor product $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$. As in the case $n=1$, we write

$$(f_1\otimes\cdots\otimes f_n)*\phi(z) = f_1(\phi(z_1))\cdots f_n(\phi(z_n))\phi'(z_1)^{\frac{\nu_1}{2}}\cdots\phi'(z_n)^{\frac{\nu_n}{2}}$$

for the group action and $\mathcal{F}_i^\phi$ for the transforms

$$(\mathcal{F}_i^\phi f)(x_1,\dots,x_n) = \langle f,(e^i_{x_1}\otimes\cdots\otimes e^i_{x_n})*\phi\rangle,\qquad i=1,2,3,\ \phi\in G.$$

Applying Proposition 4.1 in each variable gives factorizations of these transforms.

We fix a highest weight polynomial Q of degree s and write

$$\mu = |\nu| + 2s.$$

Then there is a unique intertwining embedding $\mathcal{K}_Q$ of $\mathcal{A}^\mu$ into $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ such that $\mathcal{K}_Q 1 = Q$. We will henceforth denote by $\langle\,,\,\rangle_\mu$ and $\langle\,,\,\rangle_\nu$ the scalar product of $\mathcal{A}^\mu$ and $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$, respectively.

THEOREM 4.2. *For $g\in\mathcal{A}^\mu$ and $\phi\in G$, we have the equalities*

$$\begin{aligned} (\mathcal{F}_1^\phi\mathcal{K}_Q g)(m) &= T_1Q(m)(\mathcal{F}_1^\phi g)(|m|-s),\\ (\mathcal{F}_2^\phi\mathcal{K}_Q g)(\xi) &= T_2Q(\xi)(\mathcal{F}_2^\phi g)(|\xi|),\\ (\mathcal{F}_3^\phi\mathcal{K}_Q g)(X) &= T_3Q(X)(\mathcal{F}_3^\phi g)(|X|). \end{aligned}$$

Thus we may write

$$\langle\mathcal{K}_Q g,(z_1^{m_1}\cdots z_n^{m_n})*\phi\rangle_\nu = T_1Q(m_1,\dots,m_n)\langle g, z^{|m|-s}*\phi\rangle_\mu,$$

while for T_2 and T_3 similar equalities are valid with a suitable interpretation. In particular the matrix elements

$$\langle\mathcal{K}_Q e^i_x,(e^j_{y_1}\otimes\cdots\otimes e^j_{y_n})*\phi\rangle_\nu$$

of the embedding $\mathcal{K}_Q$ factor as $T_jQ(y)$ times a matrix element (in the sense of section 3) for $\mathcal{A}^{|\nu|+2s}$.

Proof. Since $\mathcal{K}_Q$ commutes with the group action, we may assume that $\phi=\mathrm{id}$. Moreover, it suffices to choose $g=k_w$ as a reproducing kernel. The first equality then reduces to

$$\langle\mathcal{K}_Q k_w, e^1_{m_1}\otimes\cdots\otimes e^1_{m_n}\rangle_\nu = T_1Q(m)\bar w^{|m|-s}. \tag{4.2}$$

First note that

$$k_w = (1-|w|^2)^{-\frac{\mu}{2}}\, 1 * \phi_w^{-1},$$

where ϕ_w is the disc automorphism $\phi_w(z) = (z+w)/(1+\bar{w}z)$, corresponding to the matrix $\left(\begin{smallmatrix}\alpha & \beta\\ \bar{\beta} & \bar{\alpha}\end{smallmatrix}\right)$, with

$$\alpha = \frac{1}{\sqrt{1-|w|^2}}, \qquad \beta = \frac{w}{\sqrt{1-|w|^2}}.$$

This gives

$$\mathcal{K}_Q k_w = (1-|w|^2)^{-\frac{\mu}{2}}\, Q * \phi_w^{-1}.$$

Thus we have

$$\begin{aligned}\langle \mathcal{K}_Q k_w, e^1_{m_1}\otimes\cdots\otimes e^1_{m_n}\rangle_\nu &= (1-|w|^2)^{-\frac{\mu}{2}}\langle Q*\phi_w^{-1}, e^1_{m_1}\otimes\cdots\otimes e^1_{m_n}\rangle_\nu\\ &= (1-|w|^2)^{-\frac{\mu}{2}}(\mathcal{F}_1^{\phi_w}Q)(m_1,\dots,m_n).\end{aligned}$$

We now apply the factorization (4.1) to each variable. Note that (1.3) may be written as

$$(\delta_a\tau_b)^{\otimes n}Q = a^s Q.$$

This gives

$$\begin{aligned}\langle \mathcal{K}_Q k_w, e^1_{m_1}\otimes\cdots\otimes e^1_{m_n}\rangle_\nu &= (1-|w|^2)^{-\frac{\mu}{2}}\frac{\bar{\beta}^{|m|}}{\alpha^{|\nu|+|m|}}(T_1(\delta_{1/\alpha\bar{\beta}}\,\tau_{-\beta/\alpha})^{\otimes s}Q)(m)\\ &= (1-|w|^2)^{-\frac{\mu}{2}}\frac{\bar{\beta}^{|m|-s}}{\alpha^{|\nu|+|m|+s}}\,T_1Q(m) = \bar{w}^{|m|-s}T_1Q(m),\end{aligned}$$

which proves (4.2). The rest of the theorem follows from the equalities

$$\mathcal{F}_2^{\phi_w}Q(\xi) = (1-|w|^2)^{\frac{\mu}{2}}\,T_2Q(\xi)\,\overline{e^2_{|\xi|}(w)},$$

$$\mathcal{F}_3^{\phi_w}Q(X) = (1-|w|^2)^{\frac{\mu}{2}}\,T_3Q(X)\,\overline{e^3_{|X|}(w)},$$

which are likewise easily obtained from Proposition 4.1. □

Choosing as in the proof, $g = k_w$, $\phi = \mathrm{id}$, and expressing the left-hand sides by means of the adjoint $\mathcal{K}_Q^*$, one sees that the theorem is equivalent to the equalities

$$\begin{aligned}\mathcal{K}_Q^*\, e^1_{m_1}\otimes\cdots\otimes e^1_{m_n} &= \overline{T_1Q(m)}\, e^1_{|m|-s},\\ \mathcal{K}_Q^*\, e^2_{\xi_1}\otimes\cdots\otimes e^2_{\xi_n} &= \overline{T_2Q(\xi)}\, e^2_{|\xi|},\\ \mathcal{K}_Q^*\, e^3_{X_1}\otimes\cdots\otimes e^3_{X_n} &= \overline{T_3Q(X)}\, e^3_{|X|}.\end{aligned}$$

This may also be proved using the explicit expression for $\mathcal{K}_Q^*$ as a differential operator *(transvectant)* given in [R]; cf. [P2] for the case $n=2$.

We now combine Theorem 4.2 with the equalities (3.2), choosing $f = \mathcal{K}_Q e^i_x$. This will give the expansions of $\mathcal{K}_Q e^i_x$ in tensor products of the e^j_y. By the theorem, the coefficients will involve the polynomial T_jQ times the matrix element K^{ij}. For $i=j$ this exhibits the polynomials T_iQ as coupling coefficients of our representation. We summarize this in a corollary.

COROLLARY 4.3. *In the notation above, we have the generalized Clebsch–Gordan formulas*

$$\frac{(\mu)_k}{k!}\mathcal{K}_Q e^1_k(z_1,\dots,z_n) = \sum_{|m|=k+s} \frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}\, T_1Q(m) z_1^{m_1}\cdots z_n^{m_n},$$

$$\begin{aligned}&\frac{\xi^{\mu-1}}{\Gamma(\mu)}\mathcal{K}_Q e^2_\xi(z_1,\dots,z_n)\\ &= \frac{1}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\int_{\eta\in\mathbb{R}^n_+\,:\,|\eta|=\xi} T_2Q(\eta)\, e^2_{\eta_1}(z_1)\cdots e^2_{\eta_n}(z_n)\eta_1^{\nu_1-1}\cdots\eta_n^{\nu_n-1}\,d\eta,\end{aligned} \tag{4.3}$$

$$\begin{aligned}&\frac{\left|\Gamma\left(\frac{\mu}{2}+iX\right)\right|^2}{\Gamma(\mu)}\mathcal{K}_Q e^3_X(z_1,\dots,z_n) = \frac{1}{(2\pi)^{n-1}\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\\ &\times\int_{Y\in\mathbb{R}^n\,:\,|Y|=X} T_3Q(Y)\, e^3_{Y_1}(z_1)\cdots e^3_{Y_n}(z_n)\left|\Gamma\left(\frac{\nu_1}{2}+iY_1\right)\cdots\Gamma\left(\frac{\nu_n}{2}+iY_n\right)\right|^2 dY.\end{aligned}$$

To obtain the latter two equalities, one may formally use (3.3) and (3.4). This can be justified either by a duality argument or by first writing down an analogous expansion for $\mathcal{K}_Q e^i_x * \phi$ and then sending ϕ to id along a suitable one-parameter subgroup.

Choosing $i \neq j$ gives six additional expansion formulas. These follow, however, from formulas which have been obtained already. For instance, $i=1$, $j=2$ gives

$$\mathcal{K}_Q e^1_k(z) = \frac{(-1)^k}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\int_{\mathbb{R}^n_+} {}_1F_1\left(\begin{array}{c|c} -k & \\ \mu & |\eta|\end{array}\right) T_2Q(\eta)\, e^{-|\eta|}\prod_{r=1}^n e^2_{\eta_r}(z_r)\eta_r^{\nu_r-1}\,d\eta.$$

Integrating over the hypersurfaces $|\eta|=\xi$ and assuming (4.3), we see that this reduces to

$$\mathcal{K}_Q e^1_k(z) = \frac{(-1)^k}{\Gamma(\mu)}\int_0^\infty {}_1F_1\left(\begin{array}{c|c} -k & \\ \mu & \xi\end{array}\right)\mathcal{K}_Q e^2_\xi(z) e^{-\xi}\xi^{\mu-1}\,d\xi.$$

However, this follows from applying $\mathcal{K}_Q$ to each side of

$$e^1_k(z) = \frac{1}{\Gamma(\mu)}\int_0^\infty K^{12}(k,\xi,\mathrm{id})\, e^2_\xi(z) e^{-\xi}\xi^{\mu-1}\,d\xi,$$

which is a special case of (3.2).

As a consequence of the interpretation as coupling coefficients, we recover the orthogonality relations for the polynomials T_iQ with the exact proportionality constants.

COROLLARY 4.4. *Let Q and $\tilde{Q}$ be two highest weight polynomials in*

$$\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n},$$

not necessarily of the same degree. Then if s is the degree of Q (or of $\tilde{Q}$) and $\mu = |\nu|+2s$, the quantities

$$\frac{k!}{(\mu)_k}\sum_{|m|=k+s}\frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}\, T_1Q(m)\,\overline{T_1\tilde{Q}(m)},$$

$$\frac{\Gamma(\mu)B^{1-\mu}}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\int_{\xi\in\mathbb{R}^n_+\,:\,|\xi|=B} T_2Q(\xi)\,\overline{T_2\tilde{Q}(\xi)}\,\xi_1^{\nu_1-1}\cdots\xi_n^{\nu_n-1}\,d\xi,$$

$$\frac{\Gamma(\mu)}{(2\pi)^{n-1}\left|\Gamma\left(\frac{\mu}{2}+iC\right)\right|^2}\int_{X\in\mathbb{R}^n\,:\,|X|=C} T_3Q(X)\,\overline{T_3\tilde{Q}(X)}\,\prod_{r=1}^n\frac{\left|\Gamma\left(\frac{\nu_r}{2}+iX_r\right)\right|^2}{\Gamma(\nu_r)}\,dX$$

are, for each $k=0,1,2,\ldots$, $B\in\mathbb{R}_+$ *and* $C\in\mathbb{R}$, *all equal to the scalar product* $\langle Q,\tilde{Q}\rangle_\nu$.

Proof. First note that, as a consequence of Schur's lemma,

$$\langle\mathcal{K}_Qf,\mathcal{K}_{\tilde{Q}}g\rangle_\nu=\langle Q,\tilde{Q}\rangle_\nu\langle f,g\rangle_\mu,$$

where we have used $\mathcal{K}_Q1=Q$. If we put $\xi=B$ in the equality (4.3) and then take the scalar product with $\mathcal{K}_{\tilde{Q}}e_B^2*\phi$ for a suitable ϕ, the left-hand side then equals

$$\frac{B^{\mu-1}}{\Gamma(\mu)}\langle Q,\tilde{Q}\rangle_\nu\langle e_B^2,e_B^2*\phi\rangle_\mu,$$

while the right-hand side equals

$$\begin{aligned}&\frac{1}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\int_{\eta\in\mathbb{R}^n_+\,:\,|\eta|=B} T_2Q(\eta)\langle e^2_{\eta_1}\otimes\cdots\otimes e^2_{\eta_n},\mathcal{K}_{\tilde{Q}}e_B^2*\phi\rangle_\nu\,\eta_1^{\nu_1-1}\cdots\eta_n^{\nu_n-1}\,d\eta\\&=\frac{1}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\int_{\eta\in\mathbb{R}^n_+\,:\,|\eta|=B} T_2Q(\eta)\,\overline{T_2\tilde{Q}(\eta)}\langle e_B^2,e_B^2*\phi\rangle_\mu\,\eta_1^{\nu_1-1}\cdots\eta_n^{\nu_n-1}\,d\eta.\end{aligned}$$

Dividing out the matrix element $\langle e_B^2,e_B^2*\phi\rangle_\mu$ we obtain the second part of the corollary. The remaining parts are similar. □

5. Convolution formulas. In this section we will obtain some convolution formulas for the polynomials T_iQ. The most general formulas of this type would follow from considering a scalar product

$$\langle\mathcal{K}_Qe_x^i*\phi,(e^j_{y_1}\otimes\cdots\otimes e^j_{y_n})*\psi\rangle_\nu$$

and expanding this by applying formula number k in (3.1) on each variable. For instance, with $(i,j,k)=(1,2,3)$ this gives

$$\begin{aligned}&\langle\mathcal{K}_Qe_k^1*\phi,(e^2_{\xi_1}\otimes\cdots\otimes e^2_{\xi_n})*\psi\rangle_\nu\\&=\int_{\mathbb{R}^n}\langle\mathcal{K}_Qe_k^1*\phi,e^3_{X_1}\otimes\cdots\otimes e^3_{X_n}\rangle_\nu\langle e^3_{X_1}\otimes\cdots\otimes e^3_{X_n},(e^2_{\xi_1}\otimes\cdots\otimes e^2_{\xi_n})*\psi\rangle_\nu\,dm(X),\end{aligned}$$

where

$$dm(X)=\frac{1}{(2\pi)^n\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\left|\Gamma\left(\frac{\nu_1}{2}+iX_1\right)\cdots\Gamma\left(\frac{\nu_n}{2}+iX_n\right)\right|^2dX_1\cdots dX_n.$$

Now, by Theorem 4.2, this may be written as

$$T_2Q(\xi)\,K^{12}(k,|\xi|,\psi\phi^{-1})=\int_{\mathbb{R}^n}T_3Q(X)\,K^{13}(k,|X|,\phi^{-1})\prod_{r=1}^nK^{32}(X_r,\xi_r,\psi)\,dm(X).$$

Inserting the expressions for matrix elements from Proposition 3.2, we obtain an integral formula linking T_2Q and T_3Q.

We will not write down all such formulas since they may be deduced from a few special cases. Using addition formulas for the matrix elements we may reduce ourselves to the case $\phi = \mathrm{id}$. As in the previous section, we may also assume that $i = k$. In this case, the formulas follow from applying the transform $\mathcal{F}_j^{\psi}$ to both sides of the three generalized Clebsch–Gordan formulas of Corollary 4.3. Moreover, the formulas with $j = 3$ follow from those with $j = 1$, by means of (1.8). After these reductions, there remain the six cases

$$(i,j) = (1,1),\ (1,2),\ (2,1),\ (2,2),\ (3,1),\ (3,2),$$

which will give the following convolution formulas.

THEOREM 5.1. *Let, as above, Q be a highest weight polynomial in $\mathcal{A}^{\nu_1} \cdots \otimes \mathcal{A}^{\nu_n}$ of degree s and let $\mu = |\nu| + 2s$. Then the following formulas hold*:

$$\begin{aligned}(5.1)\quad & \frac{(\mu)_k}{k!} c^s \, {}_2F_1\left(\begin{matrix} -k, s-|l| \\ \mu \end{matrix} \,\middle|\, c \right) T_1 Q(l) \\ &= \sum_{|m|=k+s} \frac{(\nu_1)_{m_1} \cdots (\nu_n)_{m_n}}{m_1! \cdots m_n!} T_1 Q(m) \prod_{r=1}^{n} {}_2F_1\left(\begin{matrix} -m_r, -l_r \\ \nu_r \end{matrix} \,\middle|\, c \right), \\ & \frac{(\mu)_k}{k!} \, {}_1F_1\left(\begin{matrix} -k \\ \mu \end{matrix} \,\middle|\, |\xi| \right) T_2 Q(\xi) \\ &= (-1)^s \sum_{|m|=k+s} \frac{(\nu_1)_{m_1} \cdots (\nu_n)_{m_n}}{m_1! \cdots m_n!} T_1 Q(m) \prod_{r=1}^{n} {}_1F_1\left(\begin{matrix} -m_r \\ \nu_r \end{matrix} \,\middle|\, \xi_r \right), \end{aligned}$$

$$\begin{aligned}(5.2)\quad & \frac{\xi^{\mu-1}}{\Gamma(\mu)} \, {}_1F_1\left(\begin{matrix} s-|m| \\ \mu \end{matrix} \,\middle|\, \xi \right) T_1 Q(m) \\ &= \frac{(-1)^s}{\Gamma(\nu_1) \cdots \Gamma(\nu_n)} \int_{\eta \in \mathbb{R}_+^n \,:\, |\eta| = \xi} T_2 Q(\eta) \prod_{r=1}^{n} {}_1F_1\left(\begin{matrix} -m_r \\ \nu_r \end{matrix} \,\middle|\, \eta_r \right) \eta_1^{\nu_1 - 1} \cdots \eta_n^{\nu_n - 1} \, d\eta, \\ & \frac{\xi^{\mu-1}}{\Gamma(\mu)} \, {}_0F_1\left(\begin{matrix} - \\ \mu \end{matrix} \,\middle|\, -\xi|\zeta| \right) T_2 Q(\zeta) \\ &= \frac{(-1)^s}{\Gamma(\nu_1) \cdots \Gamma(\nu_n)} \int_{\eta \in \mathbb{R}_+^n \,:\, |\eta| = \xi} T_2 Q(\eta) \prod_{r=1}^{n} {}_0F_1\left(\begin{matrix} - \\ \nu_r \end{matrix} \,\middle|\, -\eta_r \zeta_r \right) \eta_1^{\nu_1 - 1} \cdots \eta_n^{\nu_n - 1} \, d\eta, \end{aligned}$$

$$\begin{aligned}(5.3)\quad & \frac{\left|\Gamma\left(\frac{\mu}{2} + iX\right)\right|^2}{\Gamma(\mu)} c^s \, {}_2F_1\left(\begin{matrix} s-|m|, \frac{\mu}{2} + iX \\ \mu \end{matrix} \,\middle|\, c \right) T_1 Q(m) \\ &= \frac{1}{(2\pi)^{n-1}} \int_{Y \in \mathbb{R}^n \,:\, |Y| = X} T_3 Q(Y) \prod_{r=1}^{n} \frac{\left|\Gamma\left(\frac{\nu_r}{2} + iY_r\right)\right|^2}{\Gamma(\nu_r)} \, {}_2F_1\left(\begin{matrix} -m_r, \frac{\nu_r}{2} + iY_r \\ \nu_r \end{matrix} \,\middle|\, c \right) dY, \\ & \frac{\left|\Gamma\left(\frac{\mu}{2} + iX\right)\right|^2}{\Gamma(\mu)} \, {}_1F_1\left(\begin{matrix} \frac{\mu}{2} + iX \\ \mu \end{matrix} \,\middle|\, |\xi| \right) T_2 Q(\xi) \\ &= \frac{i^s}{(2\pi)^{n-1}} \int_{Y \in \mathbb{R}^n \,:\, |Y| = X} T_3 Q(Y) \prod_{r=1}^{n} \frac{\left|\Gamma\left(\frac{\nu_r}{2} + iY_r\right)\right|^2}{\Gamma(\nu_r)} \, {}_1F_1\left(\begin{matrix} \frac{\nu_r}{2} + iY_r \\ \nu_r \end{matrix} \,\middle|\, i\xi_r \right) dY. \end{aligned}$$

The parameter c occurring in (5.1) and (5.3) depends on the choice of ψ. More precisely, in the first of these identities $c = -1/|\beta|^2$, where $\psi(z) = (\alpha z + \beta)/(\bar{\beta} z + \bar{\alpha})$,

and in the second one $c = 2/(\alpha+\bar\beta)(\bar\alpha-\beta)$. Since both sides are polynomials in c they are valid for arbitrary c. For the remaining four identities, different choices of ψ give essentially the same result.

Quite a lot of information is contained in these formulas. Many known identities follow from the three interesting special cases when (i) $Q = 1$, (ii) $n = 2$, or (iii) a numerator parameter in the hypergeometric function on the left-hand side is 0, so that this function reduces to 1. In particular, the case $n = 2$ of the first two identities occurs in [KV1] with a similar group theoretic interpretation; cf. also [Su], [VdJ].

It should be noted that if Q is constructed by binary coupling as in section 2.4, the theorem may be proved using the case $n = 2$ and induction on n. Since one may construct a basis for the space of highest weight vectors in this way, this actually gives an alternative proof of the general case.

Some degenerate cases of our convolution formulas seem interesting enough to state explicitly. First, the polynomials T_1Q (and thus T_3Q) satisfy some related identities.

COROLLARY 5.2. *In the notation of Theorem* 5.1*, the following additional formulas are valid*:

$$\frac{(\mu+|l|-s)_k}{k!}\,T_1Q(l) = \sum_{|m|=k+s} \frac{(\nu_1+l_1)_{m_1}\cdots(\nu_n+l_n)_{m_n}}{m_1!\cdots m_n!}\,T_1Q(m), \tag{5.4}$$

$$\frac{(\mu+|l|-s)_{k+s}}{(k+s)!}\,T_1Q(l) = \sum_{|m|=k+s} \frac{(\nu_1+l_1)_{m_1}\cdots(\nu_n+l_n)_{m_n}}{m_1!\cdots m_n!}\,T_1Q(m+l), \tag{5.5}$$

$$\frac{(s-|l|)_k}{k!}\,T_1Q(l) = (-1)^s \sum_{|m|=k+s} \frac{(-l_1)_{m_1}\cdots(-l_n)_{m_n}}{m_1!\cdots m_n!}\,T_1Q(m), \tag{5.6}$$

$$\frac{(s-|l|)_{k+s}}{(k+s)!}\,T_1Q(l) = \sum_{|m|=k+s} \frac{(-l_1)_{m_1}\cdots(-l_n)_{m_n}}{m_1!\cdots m_n!}\,T_1Q(l-m). \tag{5.7}$$

Proof. This follows from the convolution formula (5.1). As remarked above, it is valid for arbitrary values of c, though in the proof $c = -1/|\beta|^2$ is negative. Putting $c = 1$ and using the Chu–Vandermonde formula

$${}_2F_1\left(\begin{matrix}-m,-l\\ \nu\end{matrix}\,\middle|\,1\right) = \frac{(\nu)_{m+l}}{(\nu)_m(\nu)_l}$$

gives the equality (5.4). If we multiply both sides of (5.1) by c^{-k-s} and use the limit relation

$$\lim_{c\to\infty}\frac{1}{c^m}\,{}_2F_1\left(\begin{matrix}-m,-l\\ \nu\end{matrix}\,\middle|\,c\right) = (-1)^m\frac{(-l)_m}{(\nu)_m} \qquad (m,\,l = 0,1,2,\ldots),$$

we obtain (5.6). Incidentally, (5.6) also follows from (5.4) together with

$$T_1Q(m) = (-1)^s\,T_1Q(-\nu-m), \tag{5.8}$$

which in turn follows from (1.8) and (1.9).

If we instead multiply both sides of (5.1) by $c^{-|l|}$ and let $c\to\infty$, we get

$$(-1)^s\frac{(\mu)_k(-k)_{|l|-s}}{k!(\mu)_{|l|-s}}\,T_1Q(l) = \sum_{|m|=k+s} \frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}(-m_1)_{l_1}\cdots(-m_n)_{l_n}}{m_1!\cdots m_n!\,(\nu_1)_{l_1}\cdots(\nu_n)_{l_n}}\,T_1Q(m).$$

Replacing k by $k+|l|$ and m by $m+l$ gives (5.5). The equality (5.7) follows from (5.5) together with (5.8). ∎

We may also obtain two integral formulas for T_1Q from Theorem 5.1. In the bilinear case, the first of these is a known formula for the Hahn polynomials (see [VK, Formula 8.2.4(7)]), while we have not found the second one in the literature.

COROLLARY 5.3. *In the notation of Theorem* 5.1*, the following additional formulas are valid*:

$$\begin{aligned}
&\frac{1}{\Gamma(\mu+|m|-s)}\,T_1Q(m)\\
&\quad=\quad \frac{1}{\Gamma(\nu_1+m_1)\cdots\Gamma(\nu_n+m_n)}\int_{\eta\in\mathbb{R}^n_+\,:\,|\eta|=1} T_2Q(\eta)\,\eta_1^{m_1+\nu_1-1}\cdots\eta_n^{m_n+\nu_n-1}\,d\eta,\\
&\frac{\Gamma\left(\frac{\mu}{2}+iX\right)\Gamma\left(\frac{\mu}{2}-iX+|m|-s\right)}{\Gamma(\mu+|m|-s)}\,T_1Q(m)\\
&\quad=\quad \frac{1}{(2\pi)^{n-1}}\int_{Y\in\mathbb{R}^n\,:\,|Y|=X} T_3Q(Y)\prod_{r=1}^{m}\frac{\Gamma\left(\frac{\nu_r}{2}+iY_r\right)\Gamma\left(\frac{\nu_r}{2}-iY_r+m_r\right)}{\Gamma(\nu_r+m_r)}\,dY.
\end{aligned}$$

Proof. This follows from the convolution formulas (5.2) and (5.3), respectively. In the integral in (5.2), replace η by $\xi\eta$ so that the integration is over $\{|\eta|=1\}$. Then let $\xi\to\infty$ and use the homogeneity of T_2Q and the limit relation

$$\lim_{\xi\to\infty}\frac{1}{\xi^m}\,{}_1F_1\left(\begin{array}{c|c} -m & \\ \nu & \xi\eta\end{array}\right)=\frac{(-1)^m}{(\nu)_m}\eta^m\qquad(m=0,1,2,\ldots).$$

The other half of the corollary is the case $c=1$ of (5.3). ∎

6. Coupling kernels. In this section we introduce some polynomials in $2n$ variables which we call *coupling kernels*. Let V_s be the subspace of $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ consisting of highest weight polynomials of degree s, and let $Q_s(z,\bar w)$ be its reproducing kernel. The coupling kernels are obtained by applying one of the transforms T_i to each variable in $Q_s(z,\bar w)$. If

$$Q_s(z,\bar w)=\sum_{|t|=|u|=s}c_{t,u}\,\bar w^t z^u,$$

we define the coupling kernels P_s^{ij} as

$$P_s^{ij}(x,y)=\sum_{|t|=|u|=s}c_{t,u}\,\overline{(T_iw^t)(x)}\,(T_jz^u)(y),\qquad i,j=1,2,3.$$

Recall the decomposition

$$\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}\simeq\bigoplus_{s=0}^{\infty}\binom{n+s-2}{n-2}\mathcal{A}^{|\nu|+2s}.$$

Let Π_s denote the orthogonal projection of $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ onto the isotypic subspace equivalent to $\binom{n+s-2}{n-2}\mathcal{A}^{|\nu|+2s}$. By general Hilbert space arguments, we may express the reproducing kernel Q_s as

$$Q_s(z,\bar w)=\sum_{|t|=|u|=s}\frac{(\nu)_t(\nu)_u}{t!\,u!}\langle\Pi_sz^t,z^u\rangle_\nu\,\bar w^tz^u$$

or as

$$Q_s(z,\bar w)=\sum_{k\in\Lambda_s}\overline{Q_k(w)}\,Q_k'(z),$$

where $(Q_k)_{k\in\Lambda_s}$ is any basis of V_s and $(Q'_k)_{k\in\Lambda_s}$ is the dual basis (here Λ_s is a suitable index set of cardinality $\binom{n+s-2}{n-2}$). Thus the coupling kernels are given by

$$P_s^{ij}(x,y)=\sum_{|t|=|u|=s}\frac{(\nu)_t(\nu)_u}{t!\,u!}\langle\Pi_s z^t,z^u\rangle_\nu\overline{(T_iw^t)(x)}\,(T_jz^u)(y) \tag{6.1}$$

$$=\sum_{k\in\Lambda_s}\overline{T_iQ_k(x)}\,T_jQ'_k(y). \tag{6.2}$$

It will turn out that the coupling kernels may be expressed more explicitly in terms of Karlsson's generalized Kampé de Fériet functions (2.1). Moreover, we will recover as very special cases the dual Appell polynomials described in section 2.3.

To motivate the introduction of the coupling kernels, consider the matrix elements of the projection Π_s:

$$\langle\Pi_s(e^i_{x_1}\otimes\cdots\otimes e^i_{x_n}),(e^j_{y_1}\otimes\cdots\otimes e^j_{y_n})*\phi\rangle_\nu,\qquad i,j=1,2,3. \tag{6.3}$$

One may express Π_s as

$$\Pi_s=\sum_{k\in\Lambda_s}\mathcal{K}_{Q'_k}\mathcal{K}^*_{Q_k},$$

where (Q_k) and (Q'_k) are as above. Using Theorem 4.2 one finds that the matrix element (6.3) equals a matrix element (in the sense of section 3) for $\mathcal{A}^{|\nu|+2s}$ times the coupling kernel $P_s^{ij}(x,y)$. For instance, one has

$$\begin{aligned}
&\langle\Pi_s(e^1_{l_1}\otimes\cdots\otimes e^1_{l_n}),(e^2_{\xi_1}\otimes\cdots\otimes e^2_{\xi_n})*\phi\rangle_\nu\\
&=\sum_{k\in\Lambda_s}\langle\mathcal{K}^*_{Q_k}e^1_{l_1}\otimes\cdots\otimes e^1_{l_n},\mathcal{K}^*_{Q'_k}(e^2_{\xi_1}\otimes\cdots\otimes e^2_{\xi_n})*\phi\rangle_{|\nu|+2s}\\
&=\sum_{k\in\Lambda_s}\overline{T_1Q_k(l)}\,T_2Q'_k(\xi)\langle e^1_{|l|-s},e^2_{|\xi|}*\phi\rangle_{|\nu|+2s}\\
&=K^{12}(|l|-s,|\xi|,\phi)\,P_s^{12}(l,\xi).
\end{aligned}$$

We now give explicit expressions for the coupling kernels. It is clear that

$$P_s^{ij}(x,y)=\overline{P_s^{ji}(y,x)}, \tag{6.4}$$

so we may assume that $i\le j$. Moreover, it follows from (1.8) that

$$P_s^{i3}(x,Y)=P_s^{i1}\left(x,iY-\frac{\nu}{2}\right), \tag{6.5}$$

which leaves us with the three cases P_s^{11}, P_s^{12}, and P_s^{22}.

THEOREM 6.1. *The coupling kernels are given in terms of Karlsson's functions* (2.1) *by*

$$\begin{aligned}
P_s^{11}(k,l)&=\frac{(-1)^s(-|k|)_s(-|l|)_s}{(|\nu|+s-1)_s s!}\,F^{2:2}_{2:1}\left(\begin{matrix}|\nu|+s-1,-s\ :\ -k,-l\\ -|k|,-|l|\ :\ \nu\end{matrix}\,\middle|\,1\right),\\
P_s^{12}(k,\xi)&=\frac{(-|k|)_s|\xi|^s}{(|\nu|+s-1)_s s!}\,F^{2:1}_{1:1}\left(\begin{matrix}|\nu|+s-1,-s\ :\ -k\\ -|k|\ :\ \nu\end{matrix}\,\middle|\,\frac{\xi}{|\xi|}\right),\\
P_s^{22}(\xi,\eta)&=\frac{(-1)^s|\xi|^s|\eta|^s}{(|\nu|+s-1)_s s!}\,F^{2:0}_{0:1}\left(\begin{matrix}|\nu|+s-1,-s\ :\ -\\ -\ :\ \nu\end{matrix}\,\middle|\,\frac{\xi\eta}{|\xi||\eta|}\right).
\end{aligned}$$

The proof of this theorem will be deferred to section 8. The variable of the $F_{0:1}^{2:0}$-function should be interpreted as

$$\frac{\xi\eta}{|\xi||\eta|} = \frac{1}{(\xi_1+\cdots+\xi_n)(\eta_1+\cdots+\eta_n)}\,(\xi_1\eta_1,\dots,\xi_n\eta_n)\,.$$

We also remark that Karlsson's $F_{0:1}^{2:0}$-function is Lauricella's F_C-function.

These expressions should be compared with what one gets from choosing a particular basis (Q_k) in (6.2). This is particularly interesting when $n=2$, when the sum has only one term. In the case of P_s^{22} one gets an expression for the product of two Jacobi polynomials as an $F_{0:1}^{2:0}$-series of two variables, that is, as an Appell F_4-series. This is a classical result of Watson [W]. For P_s^{11} one gets a similar expression for the product of two Hahn polynomials due to Gasper [G1]; cf. also [G2]. For P_s^{12} one gets the identity

$$\begin{aligned}&\frac{(-1)^s s!}{(\nu_2)_s}\,P_s^{(\nu_1-1,\nu_2-1)}\left(\frac{\xi_2-\xi_1}{\xi_1+\xi_2}\right)Q_s(k_1;\nu_1-1,\nu_2-1;k_1+k_2)\\ &= F_{1:1}^{2:1}\left(\begin{matrix}\nu_1+\nu_2+s-1,-s & : & -(k_1,k_2)\\ -k_1-k_2 & : & (\nu_1,\nu_2)\end{matrix}\;\middle|\;\frac{\xi_1}{\xi_1+\xi_2},\frac{\xi_2}{\xi_1+\xi_2}\right)\end{aligned}$$

or equivalently

$$\begin{aligned}&{}_2F_1\left(\begin{matrix}-n,n+a\\ b\end{matrix}\;\middle|\;z\right){}_3F_2\left(\begin{matrix}-n,n+a,c\\ b,d\end{matrix}\;\middle|\;1\right)\\ &=\frac{(-1)^n(a-b+1)_n}{(b)_n}\,F_{1:1}^{2:1}\left(\begin{matrix}-n,n+a & : & (c,d-c)\\ d & : & (b,a-b+1)\end{matrix}\;\middle|\;z,1-z\right),\end{aligned}$$

which we have not found in the literature. For general n one may obtain generalizations of these formulas by choosing in (6.2) the basis (2.2) of V_s or a basis constructed by binary coupling as in section 2.4.

As we mentioned above, the dual Appell polynomials described in section 2.3 may be viewed as special cases of coupling kernels. This might first seem surprising, but is easily understood.

COROLLARY 6.2. *Let (Q_t) be the basis*

$$Q_t(z) = (z_1-z_n)^{t_1}\cdots(z_{n-1}-z_n)^{t_{n-1}},\qquad |t| = t_1+\cdots+t_{n-1} = s$$

of the space of highest weight polynomials in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of degree s. Let (Q'_t) be the dual basis. For $x=(x_1,\dots,x_n)$, let $\hat{x}=(x_1,\dots,x_{n-1})$. Then the polynomials $T_iQ'_t$ may be expressed in terms of Karlsson's functions as

$$\begin{aligned}T_1Q'_t(m) &= \frac{(\hat{\nu})_t(-|m|)_s}{t!\,(|\nu|+s-1)_s}\,F_{1:1}^{1:2}\left(\begin{matrix}|\nu|+s-1 & : & -t,-\hat{m}\\ -|m| & : & \hat{\nu}\end{matrix}\;\middle|\;1\right),\\ T_2Q'_t(\xi) &= \frac{(\hat{\nu})_t(-|\xi|)^s}{t!\,(|\nu|+s-1)_s}\,F_{0:1}^{1:1}\left(\begin{matrix}|\nu|+s-1 & : & -t\\ - & : & \hat{\nu}\end{matrix}\;\middle|\;\frac{\xi_1}{|\xi|},\dots,\frac{\xi_{n-1}}{|\xi|}\right),\\ T_3Q'_t(X) &= \frac{(\hat{\nu})_t(\frac{|\nu|}{2}-i|X|)_s}{t!\,(|\nu|+s-1)_s}\,F_{1:1}^{1:2}\left(\begin{matrix}|\nu|+s-1 & : & -t,\frac12\hat{\nu}-i\hat{X}\\ \frac12|\nu|-i|X| & : & \hat{\nu}\end{matrix}\;\middle|\;1\right).\end{aligned}$$

Note that the vector parameters in Karlsson's functions have dimension $n-1$ here but dimension n in Theorem 6.1.

Proof. If $u=(u_1,\dots,u_n)$, $|u|=s$, we may write

$$\Pi_s z^u = \sum_v \langle z^u, Q_v\rangle_\nu\, Q'_v = \sum_v \overline{T_1Q_v(u)}\,Q'_v.$$

Thus

$$T_i\Pi_s z^u(x) = \sum_v \overline{T_1 Q_v(u)}\, T_i Q'_v(x) = P_s^{1i}(u,x). \tag{6.6}$$

Now choose

$$u = (t,0) = (t_1,\dots,t_{n-1},0), \qquad |t| = s. \tag{6.7}$$

Then

$$\langle z^u, Q_v\rangle_\nu = \frac{t_1!\cdots t_{n-1}!}{(\nu_1)_{t_1}\cdots(\nu_{n-1})_{t_{n-1}}}\,\delta_{t,v},$$

so that in fact

$$Q'_t = \frac{(\nu_1)_{t_1}\cdots(\nu_{n-1})_{t_{n-1}}}{t_1!\cdots t_{n-1}!}\,\Pi_s z^u. \tag{6.8}$$

Combining (6.6) and (6.8) gives

$$T_i Q'_t(x) = \frac{(\nu_1)_{t_1}\cdots(\nu_{n-1})_{t_{n-1}}}{t_1!\cdots t_{n-1}!}\,P_s^{1i}(u,x), \qquad i=1,2,3.$$

It is now easy to check that the expressions for $P_s^{1i}(u,x)$ given in Theorem 6.1 simplify to those of the corollary when u is of the form (6.7). □

7. Further properties of the coupling kernels. We continue to study the formal properties of the coupling kernels. Some formulas follow from Corollary 4.4 and Theorem 5.1 by choosing highest weight vectors Q of the form

$$Q_{isx} = \sum_{k\in\Lambda_s} \overline{T_i Q_k(x)}\, Q'_k,$$

where as above $(Q_k)_{k\in\Lambda_s}$ is any basis of V_s and $(Q'_k)_{k\in\Lambda_s}$ is the dual basis. For such Q one has

$$\langle Q, Q_{isx}\rangle_\nu = T_i Q(x), \qquad Q\in V_s,$$

so Q_{isx} is the kernel of T_i, viewed as an operator on V_s. We will write down these formulas explicitly since we find them quite interesting.

Applying Corollary 4.4 with $\tilde{Q} = Q_{isx}$ gives the following identities.

PROPOSITION 7.1. *Let Q be a highest weight polynomial in $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ of degree s, and let $\mu = |\nu| + 2s$. Then, for $i=1,2,3$, $k=0,1,2,\dots$, $B\in\mathbb{R}_+$, and $C\in\mathbb{R}$, one has the equalities*

$$\begin{aligned} T_i Q(x) &= \frac{k!}{(\mu)_k}\sum_{|m|=k+s}\frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}\,T_1 Q(m)\,P_s^{1i}(x,m) \\ &= \frac{\Gamma(\mu)B^{1-\mu}}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)}\int_{\xi\in\mathbb{R}_+^n\,:\,|\xi|=B} T_2 Q(\xi)\,P_s^{2i}(x,\xi)\,\xi_1^{\nu_1-1}\cdots\xi_n^{\nu_n-1}\,d\xi \\ &= \frac{\Gamma(\mu)}{(2\pi)^{n-1}\left|\Gamma\left(\frac{\mu}{2}+iC\right)\right|^2}\int_{X\in\mathbb{R}^n\,:\,|X|=C} T_3 Q(X)\,P_s^{3i}(x,X)\prod_{r=1}^n\frac{\left|\Gamma\left(\frac{\nu_r}{2}+iX_r\right)\right|^2}{\Gamma(\nu_r)}\,dX. \end{aligned}$$

In particular, P_s^{ii} is essentially the reproducing kernel for the space of coupling coefficients T_iQ of degree s. If we recall the connection between the polynomials T_2Q and spherical harmonics, we see that the reproducing kernels for Vilenkin's polyspherical harmonics may be expressed in terms of P_s^{22} and thus in terms of Karlsson's function $F_{0:1}^{2:0}$. This is an interesting fact which we have not found in the literature.

If we choose $\tilde{Q} = Q_{isx}$ and $Q = Q_{jty}$ in Corollary 4.4, we obtain addition formulas for the coupling kernels.

PROPOSITION 7.2. *In the notation above, one has for $i, j = 1, 2, 3$ the equalities*

$$\begin{aligned} P_s^{ij}(x,y)\,\delta_{s,t} &= \frac{k!}{(\mu)_k} \sum_{|m|=k+s} \frac{(\nu_1)_{m_1}\cdots(\nu_n)_{m_n}}{m_1!\cdots m_n!}\, P_s^{i1}(x,m) P_t^{1j}(m,y) \\ &= \frac{\Gamma(\mu) B^{1-\mu}}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)} \int_{\xi\in\mathbb{R}_+^n\,:\,|\xi|=B} P_s^{i2}(x,\xi) P_t^{2j}(\xi,y)\, \xi_1^{\nu_1-1}\cdots\xi_n^{\nu_n-1}\, d\xi \\ = \frac{\Gamma(\mu)}{(2\pi)^{n-1}\left|\Gamma\left(\frac{\mu}{2}+iC\right)\right|^2} &\int_{X\in\mathbb{R}^n\,:\,|X|=C} P_s^{i3}(x,X) P_t^{3j}(X,y) \prod_{r=1}^n \frac{\left|\Gamma\left(\frac{\nu_r}{2}+iX_r\right)\right|^2}{\Gamma(\nu_r)}\, dX. \end{aligned}$$

If we choose $Q = Q_{isx}$ in Theorem 5.1, we obtain eighteen convolution formulas.

PROPOSITION 7.3. *In the notation above, one has for $i = 1, 2, 3$ the equalities*

$$\begin{aligned} &\frac{(\mu)_k}{k!} c^s\, {}_2F_1\left(\begin{matrix} -k, s-|l| \\ \mu \end{matrix}\,\middle|\, c\right) P_s^{i1}(x,l) \\ &= \sum_{|m|=k+s} \frac{(\nu)_m}{m!} P_s^{i1}(x,m) \prod_{r=1}^n {}_2F_1\left(\begin{matrix} -m_r, -l_r \\ \nu_r \end{matrix}\,\middle|\, c\right), \\ &\frac{(\mu)_k}{k!}\, {}_1F_1\left(\begin{matrix} -k \\ \mu \end{matrix}\,\middle|\, |\xi|\right) P_s^{i2}(x,\xi) \\ &= (-1)^s \sum_{|m|=k+s} \frac{(\nu)_m}{m!}\, P_s^{i1}(x,m) \prod_{r=1}^n {}_1F_1\left(\begin{matrix} -m_r \\ \nu_r \end{matrix}\,\middle|\, \xi_r\right), \\ &\frac{\xi^{\mu-1}}{\Gamma(\mu)}\, {}_1F_1\left(\begin{matrix} s-|m| \\ \mu \end{matrix}\,\middle|\, |\xi|\right) P_s^{i1}(x,m) \\ &= \frac{(-1)^s}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)} \int_{\eta\in\mathbb{R}_+^n\,:\,|\eta|=\xi} P_s^{i2}(x,\eta) \prod_{r=1}^n {}_1F_1\left(\begin{matrix} -m_r \\ \nu_r \end{matrix}\,\middle|\, \eta_r\right) \eta_1^{\nu_1-1}\cdots\eta_n^{\nu_n-1}\, d\eta, \\ &\frac{\xi^{\mu-1}}{\Gamma(\mu)}\, {}_0F_1\left(\begin{matrix} - \\ \mu \end{matrix}\,\middle|\, -\xi|\zeta|\right) P_s^{i2}(x,\zeta) \\ &= \frac{(-1)^s}{\Gamma(\nu_1)\cdots\Gamma(\nu_n)} \int_{\eta\in\mathbb{R}_+^n\,:\,|\eta|=\xi} P_s^{i2}(x,\eta) \prod_{r=1}^n {}_0F_1\left(\begin{matrix} - \\ \nu_r \end{matrix}\,\middle|\, -\eta_r\zeta_r\right) \eta_1^{\nu_1-1}\cdots\eta_n^{\nu_n-1}\, d\eta, \\ &\frac{\left|\Gamma\left(\frac{\mu}{2}+iX\right)\right|^2}{\Gamma(\mu)} c^s\, {}_2F_1\left(\begin{matrix} s-|m|, \frac{\mu}{2}+iX \\ \mu \end{matrix}\,\middle|\, c\right) P_s^{i1}(x,m) \\ &= \frac{1}{(2\pi)^{n-1}} \int_{Y\in\mathbb{R}^n\,:\,|Y|=X} P_s^{i3}(x,Y) \prod_{r=1}^n \frac{\left|\Gamma\left(\frac{\nu_r}{2}+iY_r\right)\right|^2}{\Gamma(\nu_r)}\, {}_2F_1\left(\begin{matrix} -m_r, \frac{\nu_r}{2}+iY_r \\ \nu_r \end{matrix}\,\middle|\, c\right) dY, \end{aligned}$$

$$\frac{\left|\Gamma\left(\frac{\mu}{2}+iX\right)\right|^2}{\Gamma(\mu)}\,{}_1F_1\left(\begin{matrix}\frac{\mu}{2}+iX\\ \mu\end{matrix}\,\middle|\, i|\xi|\right)P_s^{i2}(x,\xi)$$
$$=\frac{i^s}{(2\pi)^{n-1}}\int_{Y\in\mathbb{R}^n\,:\,|Y|=X}P_s^{i3}(x,Y)\prod_{r=1}^n\frac{\left|\Gamma\left(\frac{\nu_r}{2}+iY_r\right)\right|^2}{\Gamma(\nu_r)}\,{}_1F_1\left(\begin{matrix}\frac{\nu_r}{2}+iY_r\\ \nu_r\end{matrix}\,\middle|\, i\xi_r\right)dY.$$

Degenerate cases of these formulas are obtained by choosing $Q=Q_{isx}$ in Corollaries 5.2 and 5.3.

We will now give some expansions which are different in nature from those obtained so far. They are of a type known as *linearization formulas*, since they linearize a product of orthogonal functions, expressing it as a sum of functions from the same orthogonal system. Let Π_s be the orthogonal projection introduced in section 6. Then the scalar product of $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ decomposes as

$$\langle f,g\rangle_\nu=\sum_{s=0}^\infty\langle\Pi_s f,g\rangle_\nu. \tag{7.1}$$

We insert in this equality $f=e^i_{x_1}\otimes\cdots\otimes e^i_{x_n}$, $g=(e^i_{y_1}\otimes\cdots\otimes e^i_{y_n})*\phi$. Then the left-hand side is a product of matrix elements for the spaces $\mathcal{A}^{\nu_i}$, while, as we have observed, each term on the right-hand side is the product of a coupling kernel $P_s^{ij}(x,y)$ and a matrix element for $\mathcal{A}^{|\nu|+2s}$. In view of (6.4) and (6.5) it suffices to consider the three cases

$$(i,j)=(1,1),\ (1,2),\ (2,2).$$

This gives the following linearization formulas.

Proposition 7.4. *In the notation above,*

$$\prod_{r=1}^n{}_2F_1\left(\begin{matrix}-k_r,-l_r\\ \nu_r\end{matrix}\,\middle|\, c\right)=\sum_{s=0}^{\min(|k|,|l|)}P_s^{11}(k,l)\,c^s\,{}_2F_1\left(\begin{matrix}s-|k|,s-|l|\\ |\nu|+2s\end{matrix}\,\middle|\, c\right), \tag{7.2}$$
$$\prod_{r=1}^n{}_1F_1\left(\begin{matrix}-k_r\\ \nu_r\end{matrix}\,\middle|\,\xi_r\right)=\sum_{s=0}^{|k|}P_s^{12}(k,\xi)(-1)^s\,{}_1F_1\left(\begin{matrix}s-|k|\\ |\nu|+2s\end{matrix}\,\middle|\,|\xi|\right),$$
$$\prod_{r=1}^n{}_0F_1\left(\begin{matrix}-\\ \nu_r\end{matrix}\,\middle|\,-\xi_r\eta_r\right)=\sum_{s=0}^\infty P_s^{22}(\xi,\eta)\,{}_0F_1\left(\begin{matrix}-\\ |\nu|+2s\end{matrix}\,\middle|\,-|\xi||\eta|\right).$$

For $n=2$, when the coupling kernels factor as a product of two Clebsch–Gordan coefficients, the first of these formulas is the Burchnall–Chaundy formula [BC], while the third one is attributed to Bateman in [VK]. The bilinear case is also treated in [KV2].

It is not hard to prove these formulas directly or to obtain them as special cases of more general expansion formulas. For instance, one may prove that (assuming convergence)

$$\prod_{r=1}^n{}_2F_1\left(\begin{matrix}A_r,B_r\\ C_r\end{matrix}\,\middle|\,x\right)$$
$$=\sum_{s=0}^\infty\frac{(a)_s(b)_s(-1)^s}{s!(c+s-1)_s}\,F^{2:2}_{2:1}\left(\begin{matrix}c+s-1,-s\ :\ A,B\\ a,b\ :\ C\end{matrix}\,\middle|\,1\right)x^s\,{}_2F_1\left(\begin{matrix}a+s,b+s\\ c+2s\end{matrix}\,\middle|\,x\right), \tag{7.3}$$

which reduces to (7.2) for special values of the parameters and to Chaundy's formula [C]

$$
\begin{aligned}
&{}_2F_1\left(\begin{matrix}A,B\\C\end{matrix}\,\middle|\,x\right)\\
&=\sum_{s=0}^{\infty}\frac{(a)_s(b)_s(-1)^s}{s!(c+s-1)_s}\,{}_4F_3\left(\begin{matrix}c+s-1,-s,A,B\\a,b,C\end{matrix}\,\middle|\,1\right)x^s\,{}_2F_1\left(\begin{matrix}a+s,b+s\\c+2s\end{matrix}\,\middle|\,x\right)
\end{aligned}
\tag{7.4}
$$

for $n=1$. In fact, (7.3) is a special case of an even more general expansion formula due to H. M. Srivastava [S]; cf. also [SK, pp. 340–341, Formulas (251) and (254)]. (I would like to thank Professor Srivastava for pointing this out to me.) In [R], we gave a group theoretic interpretation of (7.4), where the ${}_4F_3$-series occur as Racah coefficients (or Wigner 6-j-symbols). It would be interesting to also have an interpretation of the more general formula (7.3).

We finally note some degenerate cases of (7.2). If we put $c=1$, we get

$$
\frac{(\nu_1)_{k_1+l_1}\cdots(\nu_n)_{k_n+l_n}}{(\nu_1)_{k_1}\cdots(\nu_n)_{k_n}(\nu_1)_{l_1}\cdots(\nu_n)_{l_n}}=\sum_{s=0}^{\min(|k|,|l|)}\frac{(|\nu|+2s)_{|k|+|l|-2s}}{(|\nu|+2s)_{|k|-s}(|\nu|+2s)_{|l|-s}}\,P_s^{11}(k,l).
$$

If we multiply (7.2) by $1/c^{|l|}$ and let $c\to\infty$, we get

$$
\frac{(-k_1)_{l_1}\cdots(-k_n)_{l_n}}{(\nu_1)_{l_1}\cdots(\nu_n)_{l_n}}=(-1)^s\sum_{s=0}^{|l|}\frac{(s-|k|)_{|l|-s}}{(|\nu|+2s)_{|l|-s}}\,P_s^{11}(k,l).
$$

The case $|k|=|l|$ of this equation reads

$$
\frac{k_1!\cdots k_n!}{(\nu_1)_{k_1}\cdots(\nu_n)_{k_n}}\,\delta_{k,l}=\sum_{s=0}^{|k|}\frac{(|k|-s)!}{(|\nu|+2s)_{|k|-s}}\,P_s^{11}(k,l),\qquad |k|=|l|.
$$

This may also be deduced by inserting $f=z_1^{k_1}\cdots z_n^{k_n}$, $g=z_1^{l_1}\cdots z_n^{l_n}$ in (7.1). For $n=2$, writing $P_s^{11}(k,l)$ as the product of two ${}_3F_2$-series gives the orthogonality relations for the dual Hahn polynomials (Eberlein polynomials).

8. Proof of Theorem 6.1. In this section we will prove the explicit expressions for the coupling kernels given in Theorem 6.1. The proof will be based on formula (6.1). Thus we will first compute the matrix elements

$$
\langle\Pi_s z^t, z^u\rangle_\nu,\qquad |t|=|u|=s \tag{8.1}
$$

of the projection Π_s.

We will work with the action of the universal enveloping algebra $\mathcal{U}(\mathfrak{sl}(2,\mathbb{C}))$ derived from the group action. This associative algebra has three generators E, F, and H. The action is given on $\mathcal{A}^\nu$ by the densely defined operators (cf. [R])

$$
E=-\frac{d}{dz},\qquad F=z^2\frac{d}{dz}+\nu z,\qquad H=-\left(2z\frac{d}{dz}+\nu\right)
$$

and on $\mathcal{A}^{\nu_1}\otimes\cdots\otimes\mathcal{A}^{\nu_n}$ by

$$
E=-\sum_{i=1}^n\frac{\partial}{\partial z_i},\qquad F=\sum_{i=1}^n\left(z_i^2\frac{\partial}{\partial z_i}+\nu_i z_i\right),\qquad H=-\sum_{i=1}^n\left(2z_i\frac{\partial}{\partial z_i}+\nu_i\right).
$$

These operators satisfy the structure equations

$$[EF] = H, \qquad [HE] = 2E, \qquad [HF] = -2F$$

and the reality conditions

$$E^* = -F, \qquad F^* = -E, \qquad H^* = H.$$

Let $\mathcal{H}_s$ be the subspace of $\mathcal{A}^{\nu_1} \otimes \cdots \otimes \mathcal{A}^{\nu_n}$ consisting of homogeneous polynomials of degree s. There is then an orthogonal decomposition

$$\mathcal{H}_s = V_s \oplus F\mathcal{H}_{s-1},$$

and we are interested in the orthogonal projection $\Pi_s|_{\mathcal{H}_s}$ onto the first summand.

LEMMA 8.1. *We have*

$$EF^k\big|_{\mathcal{H}_s} = F^k E - k(|\nu| + 2s + k - 1)F^{k-1}.$$

Proof. First we observe that

$$EF\big|_{\mathcal{H}_s} = (FE + H)\big|_{\mathcal{H}_s} = FE - (|\nu| + 2s)\mathrm{Id}.$$

Proceeding by induction on k, we find that

$$\begin{aligned} EF^{k+1}\big|_{\mathcal{H}_s} &= EF^k\big|_{\mathcal{H}_{s+1}} F\big|_{\mathcal{H}_s} = F^k EF - k(|\nu| + 2s + k + 1)F^k \\ &= F^k(FE - (|\nu| + 2s)) - k(|\nu| + 2s + k + 1)F^k \\ &= F^{k+1}E - (k+1)(|\nu| + 2s + k)F^k. \qquad \square \end{aligned}$$

LEMMA 8.2. *The orthogonal projection of $\mathcal{H}_s$ onto V_s is given by*

$$\sum_{k=0}^{\infty} \frac{(-1)^k}{k!(2 - |\nu| - 2s)_k} F^k E^k.$$

Note that since $E^{s+1}\big|_{\mathcal{H}_s} = 0$, the sum terminates.

Proof. We first show that the operator maps into V_s. Using the previous lemma, we find that in general

$$\begin{aligned} E\sum_{k=0}^{\infty} c_k F^k E^k\big|_{\mathcal{H}_s} &= \sum_{k=0}^{\infty} c_k EF^k\big|_{\mathcal{H}_{s-k}} E^k \\ &= \sum_{k=0}^{\infty} c_k (F^k E - k(|\nu| + 2s - k - 1)F^{k-1})E^k \\ &= \sum_{k=0}^{\infty} (c_k - c_{k+1}(k+1)(|\nu| + 2s - k - 2))F^k E^{k+1}. \end{aligned}$$

This vanishes if

$$c_{k+1} = \frac{1}{(k+1)(|\nu| + 2s - k - 2)} c_k,$$

which is indeed solved by

$$c_k = \frac{(-1)^k}{k!(2 - |\nu| - 2s)_k}.$$

We then have, for any f in $\mathcal{H}_s$,

$$f = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!(2-|\nu|-2s)_k} F^k E^k f - \sum_{k=1}^{\infty} \frac{(-1)^k}{k!(2-|\nu|-2s)_k} F^k E^k f,$$

where the first term is in V_s. Since the second term is in the image of F, it is orthogonal to the first term. This completes the proof. □

We use this lemma to compute the matrix elements (8.1).

LEMMA 8.3. *For* $|t| = |u| = s$,

$$\langle \Pi_s z^t, z^u \rangle_\nu = \frac{(-1)^s s!}{(|\nu|+s-1)_s} F_{1:1}^{1:2}\left(\begin{matrix} |\nu|+s-1 \ : \ -t, -u \\ -s \ : \ \nu \end{matrix} \,\middle|\, 1 \right). \tag{8.2}$$

Proof. Since $F^* = -E$, the previous lemma gives

$$\langle \Pi_s z^t, z^u \rangle_\nu = \sum_{k=0}^{\infty} \frac{1}{k!(2-|\nu|-2s)_k} \langle E^k z^t, E^k z^u \rangle_\nu. \tag{8.3}$$

Since E is the sum of the commuting operators $-\frac{\partial}{\partial z_j}$, it follows from the multinomial theorem that

$$E^k(z^t) = (-1)^s k! \sum_{|j|=s-k} \frac{(-t)_j}{j!} z^j$$

and thus that

$$\langle E^k z^t, E^k z^u \rangle_\nu = (k!)^2 \sum_{|j|=s-k} \frac{(-t)_j (-u)_j}{j!\,(\nu)_j}.$$

Inserting this in (8.3), we obtain

$$\begin{aligned}
\langle \Pi_s z^t, z^u \rangle_\nu &= \sum_{k=0}^{\infty} \sum_{|j|=s-k} \frac{k!\,(-t)_j (-u)_j}{(2-|\nu|-2s)_k j!\,(\nu)_j} \\
&= \sum_{j_1,\dots,j_n=0}^{\infty} \frac{(s-|j|)!\,(-t)_j (-u)_j}{(2-|\nu|-2s)_{s-|j|} j!\,(\nu)_j} \\
&= \frac{(-1)^s s!}{(|\nu|+s-1)_s} \sum_{j_1,\dots,j_n=0}^{\infty} \frac{(|\nu|+s-1)_{|j|} (-t)_j (-u)_j}{(-s)_{|j|} j!\,(\nu)_j} \\
&= \frac{(-1)^s s!}{(|\nu|+s-1)_s} F_{1:1}^{1:2}\left(\begin{matrix} |\nu|+s-1 \ : \ -t, -u \\ -s \ : \ \nu \end{matrix} \,\middle|\, 1 \right). \qquad \square
\end{aligned}$$

We are now ready to prove Theorem 6.1.

Proof of Theorem 6.1. Insert the expression (8.2) in (6.1). In the case of P_s^{12} this gives

$$\begin{aligned}
P_s^{12}(k,\xi) &= \sum_{|t|=|u|=s} \frac{(\nu)_t (\nu)_u}{t!\,u!} \langle \Pi_s z^t, z^u \rangle_\nu (-1)^s \frac{(-k)_t\, \xi^u}{(\nu)_t (\nu)_u} \\
&= \sum_{|t|=|u|=s} \frac{s!}{(|\nu|+s-1)_s} F_{1:1}^{1:2}\left(\begin{matrix} |\nu|+s-1 \ : \ -t, -u \\ -s \ : \ \nu \end{matrix} \,\middle|\, 1 \right) \frac{(-k)_t\, \xi^u}{t!\,u!} \\
&= \frac{s!}{(|\nu|+s-1)_s} \sum_{|t|=|u|=s} \sum_{j_1,\dots,j_n=0}^{\infty} \frac{(|\nu|+s-1)_{|j|} (-t)_j (-u)_j (-k)_t\, \xi^u}{(-s)_{|j|} (\nu)_j j!\, t!\, u!}.
\end{aligned}$$

Since the terms for which $t_i \le j_i$ or $u_i \le j_i$ for some i vanish, we put

$$t = j + p, \qquad u = j + q$$

and interchange the order of summation. Then

$$\frac{(-t)_j (-u)_j}{t!\, u!} = \frac{(-j-p)_j (-j-q)_j}{(j+p)!\,(j+q)!} = \frac{1}{p!\, q!}$$

so that we get

$$\begin{aligned} &P_s^{12}(k,\xi) \\ &= \frac{s!}{(|\nu|+s-1)_s} \sum_{j\,:\,|j|\le s} \frac{(|\nu|+s-1)_{|j|}}{(-s)_{|j|}(\nu)_j j!} \sum_{|p|=s-|j|} \frac{(-k)_{j+p}}{p!} \sum_{|q|=s-|j|} \frac{\xi^{j+q}}{q!}. \end{aligned} \tag{8.4}$$

Now, by the generalized Chu–Vandermonde formula,

$$\begin{aligned} \sum_{|p|=s-|j|} \frac{(-k)_{j+p}}{p!} &= (-k)_j \sum_{|p|=s-|j|} \frac{(j-k)_p}{p!} = (-k)_j \frac{(|j|-|k|)_{s-|j|}}{(s-|j|)!} \\ &= (-1)^{|j|} \frac{(-|k|)_s}{s!} \frac{(-s)_{|j|}(-k)_j}{(-|k|)_{|j|}}. \end{aligned} \tag{8.5}$$

Similarly, the multinomial theorem gives

$$\sum_{|q|=s-|j|} \frac{\xi^{j+q}}{q!} = \xi^j \frac{|\xi|^{s-|j|}}{(s-|j|)!} = (-1)^{|j|} \frac{|\xi|^s}{s!} \frac{(-s)_{|j|} \xi^j}{|\xi|^{|j|}}.$$

Inserting these expressions in (8.4) we finally obtain

$$\begin{aligned} P_s^{12}(k,\xi) &= \frac{(-|k|)_s |\xi|^s}{(|\nu|+s-1)_s s!} \sum_{j_1,\dots,j_n=0}^{\infty} \frac{(|\nu|+s-1)_{|j|}(-s)_{|j|}(-k)_j \xi^j}{(-|k|)_{|j|}(\nu)_j |\xi|^{|j|}} \\ &= \frac{(-|k|)_s |\xi|^s}{(|\nu|+s-1)_s s!}\, F_{1:1}^{2:1}\left(\begin{matrix} |\nu|+s-1, -s \\ -|k| \end{matrix} \,\middle\vert\, \begin{matrix} -k \\ \nu \end{matrix} \,\middle\vert\, \frac{\xi}{|\xi|} \right). \end{aligned}$$

The remaining two cases are similar except that for P_s^{11} one uses the generalized Chu–Vandermonde formula twice and for P_s^{22} the multinomial theorem twice. □

9. The Fock space. So far our discussion has been based on the unit disc. However, one may study analogous problems for other Hermitean symmetric spaces. In dimension 1 one has, apart from the disc, the Riemann sphere and the plane. For the Riemann sphere, the spaces $\mathcal{A}^\nu$ should be replaced by finite-dimensional spaces of polynomials. Since the group SU(2) has only one conjugacy class of one-parameter subgroups, there will only be one analogue of the transforms T_i. This transform is similar to T_1 with the parameters ν_i replaced by *negative* integers.

A more interesting case is the plane $\mathbb{C}$. We will write down the analogues of our results in this section. Since most proofs are similar to the case of the disc, we will not give the details. First we introduce the Fock space $\mathcal{F}_\alpha$ ($\alpha > 0$), consisting of entire functions for which

$$\|f\|^2 = \frac{\alpha}{\pi} \int_{\mathbb{C}} |f(z)|^2 e^{-\alpha|z|^2}\, dxdy = \sum_{k=0}^{\infty} \frac{k!}{\alpha^k} |\hat{f}(k)|^2 < \infty.$$

This space may be viewed as a limit of $\mathcal{A}^\nu$ when $\nu \to \infty$. In fact, if $\nu > 1$, the norm of $\mathcal{A}^\nu$ is given by

$$\|f\|^2 = \frac{\nu-1}{\pi} \int_{\mathbf{D}} |f(z)|^2 (1-|z|^2)^{\nu-2}\, dxdy.$$

If we replace $\mathbf{D}$ with the disc $\{|z| < R\}$ and let ν and R tend to infinity so that $\nu/R^2 \to \alpha$, the weight tends to

$$\lim_{\nu,\, R\to\infty} \left(1 - \frac{|z|^2}{R^2}\right)^{\nu-2} = e^{-\alpha|z|^2}.$$

At the level of Lie algebras, this corresponds to deforming $\mathfrak{su}(1,1)$ into the Heisenberg algebra.

On the space $\mathcal{F}_\alpha$ we introduce the Heisenberg operators

$$\mathcal{U}_h f(z) = e^{-\alpha|h|^2/2 - \alpha\bar{h}z} f(z+h), \qquad h \in \mathbb{C}.$$

They satisfy the Heisenberg relation

$$\mathcal{U}_{h_1}\mathcal{U}_{h_2} = e^{i\alpha \mathrm{Im}(h_1\bar{h}_2)} \mathcal{U}_{h_1+h_2}$$

and are unitary operators.

As an analogue of the parabolic basis for $\mathcal{A}^\nu$ we introduce the Gauss–Weierstrass functions

$$e^2_x(z) = e^{-\alpha z^2/2 + xz},$$

and write $\mathfrak{F}_2$ for the transform

$$\mathfrak{F}_2 f(x) = \langle f, e^2_x \rangle.$$

We also write $e^1_k(z) = z^k$ and

$$\mathfrak{F}_1 f(k) = \langle f, e^1_k \rangle.$$

In analogy with (3.1), and with the same interpretation, we have the Plancherel formulas

$$\begin{aligned} \langle f, g\rangle &= \sum_{k=0}^{\infty} \frac{\alpha^k}{k!}\, \mathfrak{F}_1 f(k)\, \overline{\mathfrak{F}_1 g(k)} \\ &= \frac{1}{\sqrt{2\pi\alpha}} \int_{-\infty}^{\infty} \mathfrak{F}_2 f(x)\, \overline{\mathfrak{F}_2 g(x)}\, e^{-x^2/2\alpha}\, dx. \end{aligned} \tag{9.1}$$

The matrix elements of $\mathcal{U}_\langle$ are given by

$$\langle \mathcal{U}_h e^1_k, e^1_l \rangle = (-1)^l h^k \bar{h}^l e^{-\alpha|h|^2/2}\, {}_2F_0\left(\begin{matrix} -k, -l \\ - \end{matrix} \,\middle|\, -\frac{1}{\alpha|h|^2} \right),$$

which are essentially Charlier polynomials, and by

$$\begin{aligned} \langle e^1_k, e^2_x \rangle &= \frac{x^k}{\alpha^k}\, {}_2F_0\left(\begin{matrix} -\frac{k}{2}, \frac{1}{2} - \frac{k}{2} \\ - \end{matrix} \,\middle|\, -\frac{2\alpha}{x^2} \right) \\ &= \sum_{0 \le j \le \left[\frac{k}{2}\right]} \frac{(-1)^j k!\, x^{k-2j}}{(k-2j)!\, j!\, 2^j \alpha^{k-j}}, \end{aligned}$$

which are essentially Hermite polynomials. It will be convenient to denote the latter quantity by $H_k(x,\alpha)$, so that

$$e_x^2(z)=\sum_{k=0}^{\infty}\frac{\alpha^k}{k!}H_k(x,\alpha)z^k. \tag{9.2}$$

In the standard notation for Hermite polynomials, cf. [VK], one has

$$H_k(x,\alpha)=\frac{1}{(2\alpha)^{k/2}}H_k\left(\frac{x}{\sqrt{2\alpha}}\right),\qquad \alpha>0.$$

The operators $\mathcal{U}_h$ permute the functions e_x^2, and thus the matrix elements $\langle \mathcal{U}_h e_k^1, e_x^2\rangle$ are also given by Hermite polynomials. For the same reason, the matrix elements $\langle \mathcal{U}_h e_x^2, e_y^2\rangle$ are essentially Dirac measures. We will also consider the matrix elements $\langle \delta_c e_x^2, e_y^2\rangle$, where, as above, δ_c is the dilatation $(\delta_c f)(z)=f(cz)$. Expanding the scalar product as in (9.1), we get

$$\langle \delta_c e_x^2, e_y^2\rangle=\sum_{k=0}^{\infty}\frac{\alpha^k c^k}{k!}H_k(x,\alpha)H_k(y,\alpha). \tag{9.3}$$

Mehler's formula (see [VK, Formula 9.6.8(5)]) then gives

$$\langle \delta_c e_x^2, e_y^2\rangle=\frac{1}{\sqrt{1-c^2}}\exp\left(\frac{2cxy-c^2(x^2+y^2)}{2\alpha(1-c^2)}\right),\qquad |c|<1 \tag{9.4}$$

(cf. [P1] for another computation of this scalar product).

Later we will need power series expansions of $1/e_x^2$ and $1/\langle \delta_c e_x^2, e_y^2\rangle$. Since

$$\frac{1}{e_x^2(z)}=e^{\alpha z^2/2-xz}$$

is obtained from $e_x^2(z)$ by replacing α with $-\alpha$ and z with $-z$, it follows from (9.2) that

$$\frac{1}{e_x^2(z)}=\sum_{k=0}^{\infty}\frac{\alpha^k}{k!}H_k(x,-\alpha)z^k. \tag{9.5}$$

Similarly, if in the right-hand side of (9.4) we formally replace α by $-\alpha$ and multiply by $1-c^2$, we obtain the reciprocal quantity. Manipulating (9.3) in the same way gives

$$\begin{aligned}\frac{1}{\langle \delta_c e_x^2, e_y^2\rangle}&=(1-c^2)\sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k c^k}{k!}H_k(x,-\alpha)H_k(y,-\alpha)\\&=\sum_{k=0}^{\infty}\frac{(-1)^k\alpha^k c^k}{k!}G_k(x,y,\alpha),\end{aligned} \tag{9.6}$$

where

$$\begin{aligned}G_k(x,y,\alpha)&=H_k(x,-\alpha)H_k(y,-\alpha)-\frac{k(k-1)}{\alpha^2}H_{k-2}(x,-\alpha)H_{k-2}(y,-\alpha)\\&=\sum_{0\le i,j\le\left[\frac{k}{2}\right]}\frac{k!\,(k-2)!\,(k(k-1)-4ij)x^{k-2i}y^{k-2j}}{(k-2i)!\,(k-2j)!\,i!\,j!\,2^{i+j}\alpha^{2k-i-j}}.\end{aligned} \tag{9.7}$$

Let us now consider an arbitrary tensor product

$$\mathcal{F}_{\alpha_1} \otimes \cdots \otimes \mathcal{F}_{\alpha_n}$$

of Fock spaces. It decomposes under the group action into infinitely many copies of the space $\mathcal{F}_{|\alpha|}$. To get a finer decomposition one may add to the group auxiliary operators of the form

$$f(z) \mapsto e^{im\theta} f(e^{i\theta} z),$$

where m is an integer; cf. [VK, section 8.6.1]. Temporarily writing $\mathcal{F}_{\alpha,m}$ for the corresponding representation space, we have a decomposition

$$\mathcal{F}_{\alpha_1,m_1} \otimes \cdots \otimes \mathcal{F}_{\alpha_n,m_n} = \bigoplus_{s=0}^{\infty} \binom{n+s-2}{n-2} \mathcal{F}_{|\alpha|,|m|+s}.$$

This means that the intertwining embeddings $\mathcal{F}_{|\alpha|,|m|+s} \to \mathcal{F}_{\alpha_1,m_1} \otimes \cdots \otimes \mathcal{F}_{\alpha_n,m_n}$ are required to increase homogeneity by s. If Q is an image of 1 under such an embedding, then Q must satisfy the same condition (1.3) as in the case of the disc. In the case at hand, there is a simple expression for the corresponding embedding $\mathcal{K}_Q$:

$$(\mathcal{K}_Q g)(z_1, \ldots, z_n) = Q(z_1, \ldots, z_n)\, g\left(\frac{(\alpha, z)}{|\alpha|}\right), \tag{9.8}$$

where

$$(\alpha, z) = \alpha_1 z_1 + \cdots + \alpha_n z_n.$$

From now on we write

$$\langle f, g\rangle_\alpha = \langle f, g\rangle_{\mathcal{F}_{\alpha_1} \otimes \cdots \otimes \mathcal{F}_{\alpha_n}}, \qquad \langle f, g\rangle_{|\alpha|} = \langle f, g\rangle_{\mathcal{F}_{|\alpha|}}.$$

To proceed in analogy with the case of the disc, we should now seek factorizations of the transforms

$$\mathfrak{F}_i^h = \mathfrak{F}_i \mathcal{U}_h, \quad i = 1, 2.$$

However, we get a useful factorization only in the case $i = 1$. Then we may write

$$\mathfrak{F}_1^h = M_1\, \mathfrak{T}_1\, \delta_{-1/\bar{h}}\, \tau_h \quad (h \neq 0),$$

where

$$M_1 f(k) = e^{-\alpha|h|^2/2} (-1)^k \bar{h}^k f(k)$$

and

$$\mathfrak{T}_1 z^k(m) = (-1)^k \frac{(-m)_k}{\alpha^k}.$$

Analogous to Theorem 4.2, we have

$$(\mathfrak{F}_1^h\, \mathcal{K}_Q g)(m) = \mathfrak{T}_1 Q(m) (\mathfrak{F}_1^h g)(|m| - s). \tag{9.9}$$

Thus if

$$Q(z) = \sum_{|t|=s} c_t\, z_1^{t_1} \cdots z_n^{t_n}$$

is a highest weight vector, we have a corresponding coupling coefficient

$$\mathfrak{T}_1 Q(m_1,\dots,m_n) = (-1)^s \sum_{|t|=s} c_t \, \frac{(-m_1)_{t_1}\cdots(-m_n)_{t_n}}{\alpha_1^{t_1}\cdots\alpha_n^{t_n}},$$

which is again a polynomial of degree s.

For $i=2$ we define

$$\mathfrak{T}_2 Q(x_1,\dots,x_n) = \langle Q, e_{x_1}^2 \otimes \cdots \otimes e_{x_n}^2 \rangle_\alpha,$$

where Q is a highest weight vector. There is no very simple expression for $\mathfrak{T}_2 Q$; we must be content with writing

$$\mathfrak{T}_2 Q(x_1,\dots,x_n) = \sum_{|t|=s} c_t \, H_{t_1}(x_1,\alpha_1)\cdots H_{t_n}(x_n,\alpha_n),$$

which shows that $\mathfrak{T}_2 Q$ is a polynomial of degree s. One may also show that

$$\mathfrak{T}_2 Q(x_1+b\alpha_1,\dots,x_n+b\alpha_n) = \mathfrak{T}_2 Q(x_1,\dots,x_n), \qquad b\in\mathbb{C}. \tag{9.10}$$

Note that in the case of the disc, the polynomial $T_2 Q$ is homogeneous. Thus in both cases something of the property (1.3) is preserved. As in Theorem 4.2, we have

$$(\mathfrak{F}_2^h \mathcal{K}_Q g)(x) = \mathfrak{T}_2 Q(x)(\mathfrak{F}_2^h g)(|x|) \tag{9.11}$$

and, for the dilatations,

$$(\mathfrak{F}_2 \delta_c \mathcal{K}_Q g)(x) = c^s \mathfrak{T}_2 Q(x)(\mathfrak{F}_2 \delta_c g)(|x|). \tag{9.12}$$

In the bilinear case, $n=2$, we may take $Q(z)=(z_1-z_2)^s$. Then $\mathfrak{T}_1 Q$ are essentially Krawtchouk polynomials; in the notation of [VK] we have

$$\mathfrak{T}_1 Q(m_1,m_2) = \frac{(-m_1-m_2)_s}{\alpha_2^s} \, K_s(m_1; \tfrac{\alpha_1}{\alpha_1+\alpha_2}; m_1+m_2),$$

while the polynomials $\mathfrak{T}_2 Q$ may be expressed in terms of Hermite polynomials. Indeed, we may write

$$\begin{aligned}\mathfrak{T}_2 Q(x_1,x_2) &= \sum_{k=0}^{s} \binom{s}{k} (-1)^{s-k} H_k(x_1,\alpha_1) H_{s-k}(x_2,\alpha_2) \\ &= \left(\frac{1}{\alpha_1}+\frac{1}{\alpha_2}\right)^s H_s\left(\frac{x_1}{\alpha_1}-\frac{x_2}{\alpha_2}, \frac{1}{\alpha_1}+\frac{1}{\alpha_2}\right),\end{aligned}$$

where the last equality is a special case of the addition formula for Hermite polynomials (see [VK, Formula 9.6.8(7)]), which is in turn easily derived from (9.2). Note that this expression is in agreement with (9.10). That Krawtchouk polynomials are Clebsch–Gordan coefficients for the Heisenberg algebra is well known, cf. [VK], while for the Hermite polynomial this fact, as far as we know, first occurs in [VdJ].

We now write down some properties of the coupling coefficients. First we have the Clebsch–Gordan type formulas analogous to Corollary 4.3. In this case we may use (9.8) to write them more explicitly as

$$Q(z)\frac{(\alpha,z)^k}{k!} = \sum_{|m|=k+s} \frac{\alpha_1^{m_1}\cdots\alpha_n^{m_n}}{m_1!\cdots m_n!} \, \mathfrak{T}_1 Q(m) \, z_1^{m_1}\cdots z_n^{m_n},$$

$$\begin{aligned}&Q(z)\exp\left(-\frac{(x-(\alpha,z))^2}{2|\alpha|}\right) = \frac{1}{(2\pi)^{\frac{n-1}{2}}}\sqrt{\frac{|\alpha|}{\alpha_1\cdots\alpha_n}} \\ &\times \int_{y\in\mathbb{R}^n\,:\,|y|=x} \mathfrak{T}_2 Q(y) \exp\left(-\left(\frac{(y_1-\alpha_1 z_1)^2}{2\alpha_1}+\cdots+\frac{(y_n-\alpha_n z_n)^2}{2\alpha_n}\right)\right) dy,\end{aligned}$$

where in the last equality we have written the product of the generalized basis elements and the weight of the Plancherel measures as

$$e_y^2(z)\, e^{-y^2/2\alpha} = e^{-(y-\alpha z)^2/2\alpha}.$$

Next we have orthogonality relations as in Corollary 4.4. If Q and $\tilde{Q}$ are two highest weight vectors, then their scalar product $\langle Q, \tilde{Q}\rangle_\alpha$ equals both the sum

$$\frac{k!}{|\alpha|^k} \sum_{|m|=k+s} \frac{\alpha_1^{m_1}\cdots\alpha_n^{m_n}}{m_1!\cdots m_n!}\, \mathfrak{T}_1 Q(m)\, \overline{\mathfrak{T}_1 \tilde{Q}(m)}$$

for each $k = 0, 1, 2, \ldots$ and, for each $x \in \mathbb{R}$, the integral

$$\frac{1}{(2\pi)^{\frac{n-1}{2}}} \sqrt{\frac{|\alpha|}{\alpha_1\cdots\alpha_n}} \int_{y\in\mathbb{R}^n\,:\,|y|=x} \mathfrak{T}_2 Q(y)\, \overline{\mathfrak{T}_2 \tilde{Q}(y)} \exp\left(-\left(\frac{y_1^2}{2\alpha_1} + \cdots + \frac{y_n^2}{2\alpha_n}\right)\right) dy.$$

There are also analogues of the convolution formulas of Theorem 5.1. If Q is a highest weight polynomial of degree s, we have

$$\begin{aligned}
&\frac{|\alpha|^k}{k!} c^s\, {}_2F_0\left(\begin{matrix} -k, s-|l| \\ - \end{matrix}\,\middle|\, \frac{c}{|\alpha|}\right) \mathfrak{T}_1 Q(l) \\
&= \sum_{|m|=k+s} \frac{\alpha_1^{m_1}\cdots\alpha_n^{m_n}}{m_1!\cdots m_n!}\, \mathfrak{T}_1 Q(m) \prod_{r=1}^n {}_2F_0\left(\begin{matrix} -m_r, -l_r \\ - \end{matrix}\,\middle|\, \frac{c}{\alpha_r}\right),
\end{aligned}$$

$$\begin{aligned}
&\frac{|\alpha|^k}{k!} H_k(|x|, |\alpha|)\, \mathfrak{T}_2 Q(x) \\
&= \sum_{|m|=k+s} \frac{\alpha_1^{m_1}\cdots\alpha_n^{m_n}}{m_1!\cdots m_n!}\, \mathfrak{T}_1 Q(m) \prod_{r=1}^n H_{m_r}(x_r, \alpha_r),
\end{aligned}$$

$$\begin{aligned}
&\exp\left(-\frac{x^2}{2|\alpha|}\right) H_{|m|-s}(x, |\alpha|)\, \mathfrak{T}_1 Q(m) = \frac{1}{(2\pi)^{\frac{n-1}{2}}} \sqrt{\frac{|\alpha|}{\alpha_1\cdots\alpha_n}} \\
&\quad \times \int_{y\in\mathbb{R}^n\,:\,|y|=x} \mathfrak{T}_2 Q(y) \prod_{r=1}^n H_{m_r}(y_r, \alpha_r) \exp\left(-\left(\frac{y_1^2}{2\alpha_1} + \cdots + \frac{y_n^2}{2\alpha_n}\right)\right) dy,
\end{aligned}$$

$$\begin{aligned}
&c^s \exp\left(\frac{2cx|z| - x^2 - c^2|z|^2}{2|\alpha|(1-c^2)}\right) \mathfrak{T}_2 Q(z) = \frac{1}{(2\pi(1-c^2))^{\frac{n-1}{2}}} \sqrt{\frac{|\alpha|}{\alpha_1\cdots\alpha_n}} \\
&\quad \times \int_{y\in\mathbb{R}^n\,:\,|y|=x} \mathfrak{T}_2 Q(y) \exp\left(\sum_{j=1}^n \frac{2cy_jz_j - y_j^2 - c^2z_j^2}{2\alpha_j(1-c^2)}\right) dy.
\end{aligned}$$

For $n = 2$, the first of these formulas is the addition formula for Charlier polynomials, cf. [VK, section 8.6.5], while the second one is a generalization of the addition formula for Hermite polynomials given in [VdJ] and [KV1]. As degenerate cases of the first of these formulas we obtain analogues of the equations (5.5) and (5.6):

$$\begin{aligned}
\frac{|\alpha|^{k+s}}{(k+s)!}\, \mathfrak{T}_1 Q(l) &= \sum_{|m|=k+s} \frac{\alpha_1^{m_1}\cdots\alpha_n^{m_n}}{m_1!\cdots m_n!}\, \mathfrak{T}_1 Q(m+l), \\
\frac{(s-|l|)_k}{k!}\, \mathfrak{T}_1 Q(l) &= (-1)^s \sum_{|m|=k+s} \frac{(-l_1)_{m_1}\cdots(-l_n)_{m_n}}{m_1!\cdots m_n!}\, \mathfrak{T}_1 Q(m).
\end{aligned}$$

They are proved similarly using the limit

$$\lim_{c\to\infty}\frac{1}{c^m}\,{}_2F_0\left(\begin{matrix}-m,-l\\-\end{matrix}\,\middle|\,\frac{c}{a}\right)=(-1)^m\frac{(-l)_m}{a^m}\qquad(m,\,l=0,1,2,\dots).$$

We now turn to the coupling kernels for the present situation. They are defined as in the case of the disc and have a similar group theoretic interpretation. We denote them by $\mathfrak{P}_s^{ij}$. We may reduce ourselves to the three cases

$$\mathfrak{P}_s^{11}(k,l),\qquad \mathfrak{P}_s^{12}(k,x),\qquad \mathfrak{P}_s^{22}(x,y).$$

We first give explicit expressions for the coupling kernels similar to those given in Theorem 6.1. It seems that the kernels $\mathfrak{P}_s^{12}$ and $\mathfrak{P}_s^{22}$ do not have simple expressions in terms of Karlsson's functions. Instead we give expansions involving Hermite polynomials. Our explicit expressions are

$$\mathfrak{P}_s^{11}(k,l)=\frac{(-1)^s(-|k|)_s(-|l|)_s}{|\alpha|^s s!}\,F^{1:2}_{2:0}\left(\begin{matrix}-s & : & -k,-l\\ -|k|,-|l| & : & -\end{matrix}\,\middle|\,\frac{|\alpha|}{\alpha}\right),\tag{9.13}$$

$$\begin{aligned}&\mathfrak{P}_s^{12}(k,x)\\ &=\frac{(-|k|)_s}{|\alpha|^s}\sum_{m\,:\,|m|\le s}\frac{(-1)^{|m|}|\alpha|^{|m|}(-k)_m}{(-|k|)_{|m|}m!}\sum_{|p|=s-|m|}\frac{\alpha^p}{p!}\mathbf{H}_{p+m}(x,\alpha)\qquad(9.14)\\ &=\frac{(-1)^s(-|k|)_s}{s!}\sum_{m_1,\dots,m_n=0}^{\infty}(-1)^{|m|}\frac{(-s)_{|m|}(-k)_m}{(-|k|)_{|m|}m!}H_{s-|m|}(|x|,-|\alpha|)\mathbf{H}_m(x,\alpha),\qquad(9.15)\end{aligned}$$

$$\begin{aligned}&\mathfrak{P}_s^{22}(x,y)\\ &=\frac{(-1)^s s!}{|\alpha|^s}\sum_{m\,:\,|m|\le s}\frac{|\alpha|^{|m|}\alpha^m}{(-s)_{|m|}m!}\sum_{|p|=s-|m|}\frac{\alpha^p}{p!}\mathbf{H}_{p+m}(x,\alpha)\sum_{|q|=s-|m|}\frac{\alpha^q}{q!}\mathbf{H}_{q+m}(y,\alpha)\qquad(9.16)\\ &=\frac{(-1)^s|\alpha|^s}{s!}\sum_{m_1,\dots,m_n=0}^{\infty}\frac{(-s)_{|m|}\alpha^m}{|\alpha|^{|m|}m!}G_{s-|m|}(|x|,|y|,|\alpha|)\mathbf{H}_m(x,\alpha)\mathbf{H}_m(y,\alpha),\qquad(9.17)\end{aligned}$$

where in (9.13) we write for short

$$\frac{|\alpha|}{\alpha}=\left(\frac{\alpha_1+\dots+\alpha_n}{\alpha_1},\dots,\frac{\alpha_1+\dots+\alpha_n}{\alpha_n}\right),$$

where

$$\mathbf{H}_m(x,\alpha)=H_{m_1}(x_1,\alpha_1)\cdots H_{m_n}(x_n,\alpha_n),$$

and where G is the polynomial defined in (9.7).

As for the proof, the expression (9.13) may be proved in a similar manner as in the case of the disc, using instead of Lemma 8.2 the formula

$$\Pi_s|_{\mathcal{H}_f}=\sum_{k=0}^{\infty}\frac{1}{k!|\alpha|^k}F^kE^k,$$

where Π_s is defined as the orthogonal projection of $\mathcal{F}_{\alpha_1}\otimes\cdots\otimes\mathcal{F}_{\alpha_n}$ onto the subspace spanned by all $\mathcal{K}_Q g$, where $g\in\mathcal{F}_{|\alpha|}$ and Q is a highest weight polynomial of degree s and where

$$E=-\sum_{i=1}^n\frac{\partial}{\partial z_i},\qquad F=\sum_{i=1}^n\alpha_i z_i$$

are Lie algebra operators derived from the group action. If one continues to imitate the proof of Theorem 6.1, one arrives at the expressions (9.14) and (9.16). However, in this case there is no "multinomial theorem" to sum the inner series. To prove (9.15) and (9.17), which we believe are the natural analogues of the expressions in Theorem 6.1, we use a different approach.

Proof of (9.15). Consider a tensor product $e^2_{x_1} \otimes \cdots \otimes e^2_{x_n}$. We try to decompose it as

$$e^2_{x_1} \otimes \cdots \otimes e^2_{x_n} = \sum_{s=0}^{\infty} \mathcal{K}_{Q_{sx}} g_{sx},$$

where Q_{sx} is a highest weight polynomial of degree s. Because of (9.11) we expect that one must take $g_{sx} = e^2_{|x|}$ for all s, and thus, in view of (9.8), that

$$e^2_{x_1}(z_1) \cdots e^2_{x_n}(z_n) = e^2_{|x|}\left(\frac{(\alpha, z)}{|\alpha|}\right) \sum_{s=0}^{\infty} Q_{sx}.$$

That there is such an expansion may be verified easily by checking that the function

$$\frac{e^2_{x_1}(z_1) \cdots e^2_{x_n}(z_n)}{e^2_{|x|}\left(\frac{(\alpha,z)}{|\alpha|}\right)} = \exp\left(\sum_{i=1}^{n}\left(-\frac{\alpha_i}{2} z_i^2 + x_i z_i\right) + \frac{(\alpha, z)^2}{2|\alpha|} - \frac{|x|(\alpha, z)}{|\alpha|}\right) \tag{9.18}$$

is in the kernel of E and thus that its homogeneous parts are highest weight polynomials. This may be made more transparent by rewriting the above expression as

$$\frac{e^2_{x_1}(z_1) \cdots e^2_{x_n}(z_n)}{e^2_{|x|}\left(\frac{(\alpha,z)}{|\alpha|}\right)} = \exp\left(-\frac{1}{2|\alpha|}\sum_{i<j} \alpha_i \alpha_j (z_i - z_j)^2 + 2(x_i\alpha_j - \alpha_i x_j)(z_i - z_j)\right).$$

Now, consider the matrix element

$$\langle \Pi_s e^2_x, e^1_k \rangle_\alpha = \langle \Pi_s\, e^2_{x_1} \otimes \cdots \otimes e^2_{x_n}, e^1_{k_1} \otimes \cdots \otimes e^1_{k_n} \rangle_\alpha$$

for Π_s. As for the disc, it factors as

$$\langle \Pi_s e^2_x, e^1_k \rangle_\alpha = \mathfrak{P}^{12}_s(k, x) \langle e^2_{|x|}, e^1_{|k|-s} \rangle_{|\alpha|}.$$

On the other hand, (9.9) gives

$$\langle \Pi_s e^2_x, e^1_k \rangle_\alpha = \langle \mathcal{K}_{Q^s_x} e^2_{|x|}, e^1_k \rangle_\alpha = \mathfrak{T}_1 Q_{sx}(k) \langle e^2_{|x|}, e^1_{|k|-s} \rangle_{|\alpha|}.$$

Thus we have found that

$$\mathfrak{P}^{12}_s(k, x) = \mathfrak{T}_1 Q_{sx}(k),$$

where Q_{sx} is the term of degree s in the homogeneous expansion of the function (9.18).

Now, by (9.2)

$$e^2_{x_1}(z_1) \cdots e^2_{x_n}(z_n) = \sum_{m_1, \ldots, m_n = 0}^{\infty} \frac{\alpha^m z^m}{m!} \mathbf{H}_m(x, \alpha),$$

and by (9.5)

$$\frac{1}{e^2_{|x|}\left(\frac{(\alpha,z)}{|\alpha|}\right)} = \sum_{l=0}^{\infty} \frac{(\alpha, z)^l}{l!} H_l(|x|, -|\alpha|).$$

Thus

$$Q_{sx}(z)=\sum_{m\,:\,|m|\le s}\frac{\alpha^m z^m}{m!}\mathbf{H}_m(x,\alpha)\,\frac{(\alpha,z)^{s-|m|}}{(s-|m|)!}H_{s-|m|}(|x|,-|\alpha|).$$

Expanding the power $(\alpha,z)^{s-|m|}$ gives

$$Q_{sx}(z)=\sum_{|m|+|p|=s}\frac{\alpha^{m+p}z^{m+p}}{m!\,p!}\,\mathbf{H}_m(x,\alpha)H_{s-|m|}(|x|,-|\alpha|).$$

By the definition of $\mathfrak{T}_1$, it follows that

$$\mathfrak{T}_1Q_{sx}(k)=(-1)^s\sum_{|m|+|p|=s}\frac{(-k)_{m+p}}{m!\,p!}\,\mathbf{H}_m(x,\alpha)H_{s-|m|}(|x|,-|\alpha|).$$

Summing in p, we have as in (8.5) that

$$\sum_{|p|=s-|m|}\frac{(-k)_{m+p}}{p!}=(-1)^{|m|}\frac{(-|k|)_s}{s!}\frac{(-s)_{|m|}(-k)_m}{(-|k|)_{|m|}}.$$

Inserting this in the previous formula gives the desired expression (9.15). □

Proof of (9.17). We consider the matrix element

$$\langle\delta_c e_x^2,e_y^2\rangle_\alpha=\langle(\delta_c)^{\otimes n}e_{x_1}^2\otimes\cdots\otimes e_{x_n}^2,e_{y_1}^2\otimes\cdots\otimes e_{y_n}^2\rangle_\alpha$$

of δ_c, $|c|<1$, and expand it as

$$\langle\delta_c e_x^2,e_y^2\rangle_\alpha=\sum_{s=0}^{\infty}\langle\Pi_s\delta_c e_x^2,e_y^2\rangle_\alpha.$$

By (9.12)

$$\langle\Pi_s\delta_c e_x^2,e_y^2\rangle_\alpha=c^s\,\mathfrak{P}_s^{22}(x,y)\langle\delta_c e_{|x|}^2,e_{|y|}^2\rangle_{|\alpha|}.$$

It follows that $\mathfrak{P}_s^{22}(x,y)$ is the coefficient of c^s in

$$\frac{\langle\delta_c e_x^2,e_y^2\rangle_\alpha}{\langle\delta_c e_{|x|}^2,e_{|y|}^2\rangle_{|\alpha|}}.$$

Using the expansions (9.2) and (9.6) one obtains the expression (9.17). □

For $n=2$, the coupling kernels factor as a product of two Clebsch–Gordan coefficients. Choosing $Q(z)=(z_1-z_2)^s$, we may write

$$\mathfrak{P}_s^{ij}((x_1,x_2),(y_1,y_2))=\frac{1}{\|Q\|_\alpha^2}\,\mathfrak{T}_iQ(x_1,x_2)\,\mathfrak{T}_jQ(y_1,y_2),\qquad i,\,j=1,2,$$

where the norm is easily computed;

$$\|Q\|_\alpha^2=s!\left(\frac{1}{\alpha_1}+\frac{1}{\alpha_2}\right)^s.$$

Comparing with the expressions for $\mathfrak{P}_s^{ij}$ given above leads to Watson-type formulas, which we have not found in the literature. The case $i=j=1$ gives the identity

$$\begin{aligned}&\frac{\alpha_1^s}{\alpha_2^s}K_s(k_1;\tfrac{\alpha_1}{\alpha_1+\alpha_2};k_1+k_2)\,K_s(l_1;\tfrac{\alpha_1}{\alpha_1+\alpha_2};l_1+l_2)\\&=(-1)^s\,F^{1:2}_{2:0}\left(\begin{matrix}-s & : & -(k_1,k_2),-(l_1,l_2)\\ -k_1-k_2,-l_1-l_2 & : & -\end{matrix}\,\middle|\,\frac{\alpha_1+\alpha_2}{\alpha_1},\frac{\alpha_1+\alpha_2}{\alpha_2}\right).\end{aligned}$$

Changing the names of the variables, this may be rewritten as

$$
\begin{aligned}
&{}_2F_1\left(\begin{matrix}-n,a\\ b\end{matrix}\,\middle|\, x\right){}_2F_1\left(\begin{matrix}-n,c\\ d\end{matrix}\,\middle|\, x\right)\\
&=(1-x)^n\,F^{1:2}_{2:0}\left(\begin{matrix}-n\ :\ (a,b-a),(c,d-c)\\ b,d\ :\ -\end{matrix}\,\middle|\, x,\frac{x}{x-1}\right),
\end{aligned}
$$

which is true for arbitrary a,b,c,d since the series terminate. The case $i=1$, $j=2$ gives

$$
\begin{aligned}
&\frac{1}{\alpha_2^s}\,K_s(k_1;\tfrac{\alpha_1}{\alpha_1+\alpha_2};k_1+k_2)H_s\left(\frac{x_1}{\alpha_1}-\frac{x_2}{\alpha_2},\frac{1}{\alpha_1}+\frac{1}{\alpha_2}\right)\\
&=(-1)^s\sum_{m_1,m_2=0}^{\infty}(-1)^{m_1+m_2}\frac{(-s)_{m_1+m_2}(-k_1)_{m_1}(-k_2)_{m_2}}{(-k_1-k_2)_{m_1+m_2}m_1!\,m_2!}\\
&\quad\times H_{s-m_1-m_2}(x_1+x_2,-\alpha_1-\alpha_2)H_{m_1}(x_1,\alpha_1)H_{m_2}(x_2,\alpha_2).
\end{aligned}
$$

Finally, the case $i=j=2$ gives

$$
\begin{aligned}
&H_s\left(\frac{x_1}{\alpha_1}-\frac{x_2}{\alpha_2},\frac{1}{\alpha_1}+\frac{1}{\alpha_2}\right)H_s\left(\frac{y_1}{\alpha_1}-\frac{y_2}{\alpha_2},\frac{1}{\alpha_1}+\frac{1}{\alpha_2}\right)\\
&=(-\alpha_1\alpha_2)^s\sum_{m_1,m_2=0}^{\infty}\frac{(-s)_{m_1+m_2}\alpha_1^{m_1}\alpha_2^{m_2}}{(\alpha_1+\alpha_2)^{m_1+m_2}m_1!\,m_2!}\,G_{s-m_1-m_2}(x_1+x_2,y_1+y_2,\alpha_1+\alpha_2)\\
&\quad\times H_{m_1}(x_1,\alpha_1)H_{m_2}(x_2,\alpha_2)H_{m_1}(y_1,\alpha_1)H_{m_2}(y_2,\alpha_2).
\end{aligned}
$$

We now discuss the biorthogonal polynomial systems analogous to the Appell and dual Appell polynomials of sections 2.2 and 2.3. Thus we introduce the basis

$$
Q_t(z)=(z_1-z_n)^{t_1}\cdots(z_{n-1}-z_n)^{t_{n-1}},\qquad t_1+\cdots+t_{n-1}=s
$$

of the space of highest weight vectors of degree s in $\mathcal{F}_{\alpha_1}\otimes\cdots\otimes\mathcal{F}_{\alpha_n}$ and let (Q'_t) denote the dual basis. As in the case of the disc, it is easy to find expressions for the polynomials $\mathfrak{T}_iQ_t$ using the definition of $\mathfrak{T}_i$ directly, while the polynomials $\mathfrak{T}_iQ'_t$ arise as special cases of the coupling kernels; in fact

$$
\mathfrak{T}_iQ'_t(x)=\frac{\alpha_1^{t_1}\cdots\alpha_{n-1}^{t_{n-1}}}{t_1!\cdots t_{n-1}!}\,\mathfrak{P}^{1i}_s((t_1,\dots,t_{n-1},0),(x_1,\dots,x_n)).
$$

Writing as above $\hat{x}=(x_1,\dots,x_{n-1})$, where $x=(x_1,\dots,x_n)$, we have

$$
\begin{aligned}
\mathfrak{T}_1Q_t(m)&=\frac{(-m_n)_s}{\alpha_n^s}\,F^{0:2}_{1:0}\left(\begin{matrix}-\ :\ -t,-\hat{m}\\ 1+m_n-s\ :\ -\end{matrix}\,\middle|\,-\frac{\alpha_n}{\alpha_1},\dots,-\frac{\alpha_n}{\alpha_{n-1}}\right),\\
\mathfrak{T}_1Q'_t(m)&=\frac{\hat{\alpha}^t(-|m|)_s}{t!\,|\alpha|^s}\,F^{0:2}_{1:0}\left(\begin{matrix}-\ :\ -t,-\hat{m}\\ -|m|\ :\ -\end{matrix}\,\middle|\,\frac{|\alpha|}{\alpha_1},\dots,\frac{|\alpha|}{\alpha_{n-1}}\right),\\
\mathfrak{T}_2Q_t(x)&=(-1)^s\sum_{k_1,\dots,k_{n-1}=0}^{\infty}\frac{(-t)_k}{k!}\prod_{j=1}^{n-1}H_{k_j}(x_j,\alpha_j)H_{s-|k|}(x_n,\alpha_n),\\
\mathfrak{T}_2Q'_t(x)&=\frac{\hat{\alpha}^t}{t!}\sum_{k_1,\dots,k_{n-1}=0}^{\infty}(-1)^{|k|}\frac{(-t)_k}{k!}\prod_{j=1}^{n-1}H_{k_j}(x_j,\alpha_j)H_{s-|k|}(|x|,-|\alpha|)
\end{aligned}
$$

(Karlsson's $F^{0:2}_{1:0}$-function equals Lauricella's F_B-function). The first of these biorthogonal systems is Tratnik's multivariable Krawtchouk polynomials [T3]. We have not found the second one in the literature.

Continuing this exposition of the Fock space analogues of our results, we should now proceed with section 7. Special choices of Q in the orthogonality and convolution formulas above lead to analogues of Propositions 7.1, 7.2, and 7.3. The details of this are left to the reader. There are also analogues of the linearization formulas of Proposition 7.4. For $\mathfrak{P}_s^{11}$ we have the linearization formula

$$\prod_{r=1}^{n} {}_2F_0\left(\begin{matrix}-k_r,-l_r\\-\end{matrix}\,\middle|\,\frac{c}{\alpha_r}\right)=\sum_{s=0}^{\min(|k|,|l|)}\mathfrak{P}_s^{11}(k,l)\,c^s\,{}_2F_0\left(\begin{matrix}s-|k|,s-|l|\\-\end{matrix}\,\middle|\,\frac{c}{|\alpha|}\right).$$

Degenerate cases of this formula are

$$\frac{(-k_1)_{l_1}\cdots(-k_n)_{l_n}}{\alpha_1^{l_1}\cdots\alpha_n^{l_n}}=(-1)^s\sum_{s=0}^{|l|}\frac{(s-|k|)_{|l|-s}}{|\alpha|^{|l|-s}}\,\mathfrak{P}_s^{11}(k,l),$$

$$\frac{k_1!\cdots k_n!}{\alpha_1^{k_1}\cdots\alpha_n^{k_n}}\,\delta_{k,l}=\sum_{s=0}^{|k|}\frac{(|k|-s)!}{|\alpha|^{|k|-s}}\,\mathfrak{P}_s^{11}(k,l),\qquad |k|=|l|.$$

For $\mathfrak{P}_s^{12}$ and $\mathfrak{P}_s^{22}$ the linearization formulas seem to be of a more trivial nature. For $\mathfrak{P}_s^{12}$ we obtain

$$\prod_{r=1}^{n}H_{k_r}(x_r,\alpha_r)=\sum_{s=0}^{|k|}\mathfrak{P}_s^{12}(k,x)H_{|k|-s}(|x|,|\alpha|).$$

This is, however, very easy to deduce from (9.15). For $\mathfrak{P}_s^{22}$, the linearization formula is

$$\prod_{r=1}^{n}\langle\delta_c e_{x_r},e_{y_r}\rangle_{\alpha_r}=\sum_{s=0}^{\infty}c^s\,\mathfrak{P}_s^{22}(x,y)\langle\delta_c e_{|x|},e_{|y|}\rangle_{|\alpha|},$$

which we have already used in the proof of (9.17). As for the disc, when $n=2$ we may factor $\mathfrak{P}_s^{ij}$ as a product of two Clebsch–Gordan coefficients and obtain formulas of Burchnall–Chaundy-type. In the case of $\mathfrak{P}_s^{11}$ we obtain the multiplication formula for Charlier polynomials in this way; cf. [VK, section 8.6.5].

Acknowledgments. I am grateful to Jaak Peetre for his careful reading of this manuscript, leading to many improvements. I would also like to thank the referees for several useful comments.

REFERENCES

[A1] P. Appell, *Sur des polynômes de deux variables analogues aux polynômes de Jacobi*, Arch. Math. Phys., 66 (1881), pp. 238–245.

[A2] P. Appell, *Sur les fonctions hypergéométriques de deux variables*, J. Math. Pure Appl. (3), VIII (1882), pp. 173–216.

[AK] P. Appell and J. Kampé de Fériet, *Fonctions hypergéométriques et hypersphériques. Polynomes d'Hermite*, Gauthier–Villars, Paris, 1926.

[BW1] D. Basu and K. B. Wolf, *The unitary irreducible representations of $SL(2,\mathbb{R})$ in all subgroup reductions*, J. Math. Phys., 23 (1982), pp. 189–205.

[BW2] D. Basu and K. B. Wolf, *The Clebsch–Gordan coefficients of the three-dimensional Lorentz algebra in the parabolic basis*, J. Math. Phys., 24 (1983), pp. 478–500.

[BC] J. C. Burchnall and T. W. Chaundy, *The hypergeometric identities of Cayley, Orr, and Bailey*, Proc. London Math. Soc., 50 (1949), pp. 56–74.

[C] T. W. Chaundy, *Expansions of hypergeometric functions*, Quart. J. Math. Oxford Ser., 13 (1942), pp. 159–171.

[E] G. K. Engelis, *On polynomials orthogonal on a triangle*, Latvi. Univ. Zinat. Raksti, 58 (1964), pp. 43–48 (in Russian).

[Ex] H. Exton, *Generating relations for Tratnik's multivariable biorthogonal continuous Hahn polynomials*, J. Math. Phys., 33 (1992), pp. 524–527.

[FL] E. D. Fackerell and R. A. Littler, *Polynomials biorthogonal to Appell's polynomials*, Bull. Austral. Math. Soc., 11 (1974), pp. 181–195.

[G1] G. Gasper, *Non-negativity of a discrete Poisson kernel for the Hahn polynomials*, J. Math. Anal. Appl., 42 (1973), pp. 438–451.

[G2] G. Gasper, *Products of terminating ${}_3F_2(1)$ series*, Pacific J. Math., 56 (1975), pp. 87–95.

[KMT] E. G. Kalnins, W. Miller, Jr., and M. V. Tratnik, *Families of orthogonal and biorthogonal polynomials on the n-sphere*, SIAM J. Math. Anal., 22 (1991), pp. 272–294.

[KM1] S. Karlin and J. McGregor, *On some stochastic models in genetics*, in Stochastic Models in Medicine and Biology, J. Gurland, ed., University of Wisconsin Press, Madison, 1964, pp. 245–271.

[KM2] S. Karlin and J. McGregor, *Linear growth models with many types and multidimensional Hahn polynomials*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 261–288.

[Ka] P. W. Karlsson, *Reduction of certain generalized Kampé de Fériet functions*, Math. Scand., 32 (1973), pp. 265–268.

[KV1] H. T. Koelink and J. Van der Jeugt, *Convolutions for orthogonal polynomials from Lie and quantum algebra representations*, SIAM J. Math. Anal., 29 (1998), pp. 794–822.

[KV2] H. T. Koelink and J. Van der Jeugt, *Bilinear generating functions for orthogonal polynomials*, preprint, Universiteit van Amsterdam, Amsterdam, 1997, available at http://xxx.lanl.gov/abs/q-alg/9704016.

[K1] T. H. Koornwinder, *Two-variable analogues of the classical orthogonal polynomials*, in Theory and Application of Special Functions, R. Askey, ed., Academic Press, New York, 1975, pp. 435–495.

[K2] T. H. Koornwinder, *Group theoretic interpretation of Askey's scheme of hypergeometric orthogonal polynomials*, in Orthogonal Polynomials and Their Applications, M. Alfaro et al., eds., Lecture Notes Math. 1329, Springer-Verlag, Berlin, 1988, pp. 46–72.

[KS] T. H. Koornwinder and A. L. Schwartz, *Product formulas and associated hypergroups for orthogonal polynomials on the simplex and the parabolic biangle*, Constr. Approx., 13 (1997), pp. 537–567.

[LT] C. S. Lam and M. V. Tratnik, *Conformally invariant operator-product expansions of any number of operators of arbitrary spin*, Canad. J. Phys., 63 (1985), pp. 1427–1437.

[Me] J. Meixner, *Unformung gewisser Reihen, deren Glieder Produkte hypergeometrische Funktionen sind*, Deutsche Math., 6 (1941), pp. 341–349.

[MR] N. Mukunda and B. Radhakrishnan, *Clebsch–Gordan problem and coefficients for the three-dimensional Lorentz group in a continuous basis* I, J. Math. Phys., 15 (1974), pp. 1320–1331.

[M] G. Munschy, *Résolution de l'équation de Schrödinger des atomes à deux électrons.* III. *Suite de la méthode. Etats S symétriques*, J. Phys. Radium, 18 (1957), pp. 552–558.

[MP] G. Munschy and P. Pluvinage, *Résolution de l'équation de Schrödinger des atomes à deux électrons.* II. *Méthode rigoreuse. Etats S symétriques*, J. Phys. Radium, 18 (1957), pp. 157–160.

[P1] J. Peetre, *Some calculations related to Fock space and the Shale–Weil representation*, Integral Equations Operator Theory, 12 (1989), pp. 67–81.

[P2] J. Peetre, *Orthogonal polynomials arising in connection with Hankel forms of higher weight*, Bull. Sci. Math. (2), 116 (1992), pp. 265–284.

[Pr] J. Proriol, *Sur une famille de polynomes à deux variables orthogonaux dans un triangle*, C. R. Acad. Sci. Paris, 245 (1957), pp. 2459–2461.

[Ra] M. Rahman, *Discrete orthogonal systems corresponding to Dirichlet distribution*, Utilitas Math., 20 (1981), pp. 261–272.

[R] H. Rosengren, *Multilinear Hankel forms of higher order and orthogonal polynomials*, Math. Scand., to appear.

[Sa] P. Sally, *Analytic continuation of the irreducible unitary representations of the universal covering group of* $\mathrm{SL}(2,\mathbb{R})$, Mem. Amer. Math. Soc., 69 (1967).

[S] H. M. Srivastava, *Some polynomial expansions for functions of several variables*, IMA J. Appl. Math., 27 (1981), pp. 299–306.

[SK] H. M. Srivastava and P. W. Karlsson, *Multiple Gaussian Hypergeometric Series*, Ellis Horwood, Chichester, 1985.

[SM] H. M. Srivastava and H. L. Manocha, *A Treatise on Generating Functions*, Ellis Horwood, Chichester, 1984.

[Su] S. K. Suslov, *The Hahn polynomials in the Coulomb problem*, Sov. J. Nucl. Phys., 40 (1984), pp. 79–82 (in English); Yad. Fiz., 40 (1984), pp. 126–132 (in Russian).

[T1] M. V. Tratnik, *Multivariable continuous Hahn polynomials*, J. Math. Phys., 29 (1988), pp. 1529–1534.

[T2] M. V. Tratnik, *Multivariable biorthogonal Hahn polynomials*, J. Math. Phys., 30 (1989), pp. 627–634.

[T3] M. V. Tratnik, *Multivariable Meixner, Krawtchouk, and Meixner–Pollaczek polynomials*, J. Math. Phys., 30 (1989), pp. 2740–2749.

[T4] M. V. Tratnik, *Some multivariable orthogonal polynomials of the Askey tableau — continuous families*, J. Math. Phys., 32 (1991), pp. 2065–2073.

[T5] M. V. Tratnik, *Some multivariable orthogonal polynomials of the Askey tableau — discrete families*, J. Math. Phys., 32 (1991), pp. 2337–2342.

[VdJ] J. Van der Jeugt, *Coupling coefficients for Lie algebra representations and addition formulas for special functions*, J. Math. Phys., 38 (1997), pp. 2728–2740.

[V] N. Ya. Vilenkin, *Polyspherical and orispherical functions*, Mat. Sb., 68 (1965), pp. 432–443 (in Russian).

[VK] N. Ya. Vilenkin and A. U. Klimyk, *Representation of Lie Groups and Special Functions*, Vols. 1–3, Kluwer Academic Publishers, Dordrecht, 1991, 1992, 1993.

[W] G. N. Watson, *The product of two hypergeometric functions*, Proc. London Math. Soc., 20 (1922), pp. 189–195.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 273–297

THE EVANS FUNCTION AND GENERALIZED MELNIKOV INTEGRALS*

TODD KAPITULA†

Abstract. The Evans function, $E(\lambda)$, is an analytic function whose zeros coincide with the eigenvalues of the operator, L, obtained by linearizing about a travelling wave. The algebraic multiplicity of the eigenvalue λ_0 is equal to the order of the zero of $E(\lambda)$. If m is the geometric multiplicity and p is the algebraic multiplicity of the eigenvalue, the term $\partial_\lambda^p E(\lambda_0)$ is shown to be proportional to the determinant of an $m \times m$ matrix whose entries are given by the L^2 inner products of the eigenfunctions of the adjoint operator L^* and the generalized eigenfunctions of L. Perturbation expressions are then derived for coefficients in the Taylor expansion of $E(\lambda)$ at $\lambda = \lambda_0$ in the circumstance that the algebraic multiplicity of the eigenvalue decreases under perturbation. The expressions are used to study the eigenvalue structure for operators obtained by linearizing about bright solitary wave solutions to perturbed nonlinear Schrödinger equations.

Key words. travelling waves, stability, Evans function

AMS subject classifications. 34A26, 34C35, 34C37, 35K57, 35P15

PII. S0036141097327963

1. Introduction.

Consider the semilinear parabolic system

$$u_t = Bu_{xx} + f(u, u_x), \tag{1.1}$$

where $(x,t) \in \mathbf{R} \times \mathbf{R}^+$, B is a positive semidefinite invertible $n \times n$ matrix, $u \in \mathbf{R}^n$, and the nonlinearity f is at least C^3. A travelling wave solution $\phi(z)$ is a C^2 function of the variable $z = x - ct$ which is a solution to (1.1) and satisfies

$$\lim_{z \to \pm\infty} \phi(z) = \phi_\pm, \quad f(\phi_\pm, 0) = 0.$$

The approach to the constant states is assumed to be exponentially fast. For the rest of this paper, assume that the wave exists when $c = c^*$, so that the wave is then a time-independent solution to

$$u_t = Bu_{zz} + c^* u_z + f(u, u_z). \tag{1.2}$$

Once a wave has been shown to exist, one would then like to determine its stability relative to small perturbations. When attempting to determine the stability of the wave, it is natural to study the spectrum of the linear operator L, where

$$L = B\partial_z^2 + P(z)\partial_z + N(z), \tag{1.3}$$

and

$$P(z) = c^* + Df_{u_z}(\phi, \phi_z), \quad N(z) = Df_u(\phi, \phi_z).$$

*Received by the editors September 29, 1997; accepted for publication (in revised form) April 29, 1998; published electronically November 17, 1998. This work was partially supported by the AFOSR under grant F49620-94-1-0007.

http://www.siam.org/journals/sima/30-2/32796.html

†Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131 (kapitula@math.unm.edu).

Note that the assumption on the wave implies that

$$\lim_{z\to\pm\infty} N(z) = N_\pm, \quad \lim_{z\to\pm\infty} P(z) = P_\pm,$$

with the approach being exponentially fast. By the assumption on the matrix B, L generates a C^0-semigroup [26]. The essential spectrum of L, hereby denoted $\sigma_e(L)$, is bounded by the curves

$$\Gamma_\pm = \{\lambda : |-B\tau^2 + iP_\pm\tau + N_\pm - \lambda I_n| = 0, \tau \in \mathbf{R}\}$$

(see [11]). Set

$$\Omega = \mathbf{C}\backslash\sigma_e(L). \tag{1.4}$$

The only spectrum of L in Ω is the point spectrum, i.e., isolated eigenvalues of finite multiplicity. If there exists a $\delta > 0$ such that $\sigma_e(L) \subset \{\lambda : \operatorname{Re}\lambda \leq \delta\}$, the constant solutions $u = \phi_\pm$ are stable solutions to (1.2) [11]. If the constant solutions are stable, the stability of the wave is determined by the location of the point spectrum. If there exists an eigenvalue λ_0 with $\operatorname{Re}\lambda_0 > 0$, then the wave is exponentially unstable; otherwise, the wave is marginally stable. If the only eigenvalue with $\operatorname{Re}\lambda \geq 0$ is the one at $\lambda = 0$, and if this eigenvalue is semisimple, then the wave is exponentially orbitally stable [6], [11].

The Evans function, $E(\lambda)$, is an analytic function for $\lambda \in \Omega$ with the property that the zeros coincide with the eigenvalues of L; furthermore, the order of the zero is the algebraic multiplicity (a.m.) of the eigenvalue ([1], [29], and the references contained therein). Once the location of an eigenvalue has been determined, i.e., once a $\lambda_0 \in \Omega$ has been found such that $E(\lambda_0) = 0$, one usually wishes to determine $\partial_\lambda^p E(\lambda_0)$, where p is the order of the zero. This is especially the case if $\lambda_0 = 0$ and one wishes to determine the location of small eigenvalues near zero. Examples of such a calculation are in a plethora of papers [1], [2], [3], [4], [7], [8], [9], [12], [15], [27], [28], [29], [30], [31], [33], so that it is clearly of interest to make such a computation.

Solutions to (1.2) are invariant under spatial translation, so that $L\phi_z = 0$; thus, $\lambda_0 = 0$ is an eigenvalue of L. If (1.2) possesses no other symmetry, then the eigenvalue $\lambda_0 = 0$ is generically simple, so that $E(0) = 0$ with $\partial_\lambda E(0) \neq 0$. Alexander and Jones [3], [4] developed the orientation index for determining the sign of $\partial_\lambda E(0)$, and showed that the index is related to the manner in which the wave is constructed in the ODE phase space. Rubin [32], Rubin and Jones [33], and Sandstede [34] have shown that the orientation index is equal to the formulation of $\partial_\lambda E(0)$ given in this paper. A limitation of the index is that if $\partial_\lambda E(0) = 0$, then one is unable to use it to determine if an eigenvalue is moving through the origin into the right- or left-half plane, unless it is known a priori that there is only one other eigenvalue near $\lambda = 0$. Thus, at a minimum it is of interest to find a calculable way to determine $\partial_\lambda^2 E(0)$.

Let $\lambda_0 \in \Omega$ be an eigenvalue with geometric multiplicity (g.m.) $= m$ and a.m. $= p$. Following Gardner and Jones [8] (see Lemma 2.1), the generalized eigenfunctions $\psi_{j,i}$, $i = 1, \ldots, m$, $j = 1, \ldots, a_i$, with $p = \sum_{i=1}^m a_i$, can be ordered such that

$$(L - \lambda_0)\psi_{j,i} = \psi_{j-1,i}, \quad \psi_{0,i} = 0. \tag{1.5}$$

Since g.m.$= m$, there exist adjoint solutions $v_i \in \mathcal{N}((L-\lambda_0)^*)$, $i = 1, \ldots, m$. In all that follows, let $\langle\cdot,\cdot\rangle$ represent the L^2 inner product of complex vector-valued functions.

THEOREM 1.1. *Suppose that λ_0 is an isolated eigenvalue with g.m. $= m$ and a.m. $= p$. Let the functions $\psi_{j,i}$ be as defined in* (1.5)*, and let v_i represent the adjoint eigenfunctions. Then*

$$\partial_\lambda^p E(\lambda_0) = (-1)^{m_s} p! \begin{vmatrix} \langle \psi_{a_1,1}, v_1 \rangle & \cdots & \langle \psi_{a_1,1}, v_m \rangle \\ \vdots & & \vdots \\ \langle \psi_{a_m,m}, v_1 \rangle & \cdots & \langle \psi_{a_m,m}, v_m \rangle \end{vmatrix},$$

where $m_s = \sum_{i=0}^{m-1} i$.

Remark 1.2. As is done in Gardner and Jones [8], the above result can be extended to the case where the differential operator has order higher than two.

For the moment, assume that $\lambda_0 = 0$ and g.m. = a.m. = 1. Then $\psi_{1,1} = \phi_z$, and as a consequence of the above theorem,

$$\partial_\lambda E(0) = \langle \phi_z, v \rangle,$$

where v is the unique (up to scalar multiplication) adjoint eigenfunction of L. As is discussed in Alexander and Jones [3], [4], $\partial_\lambda E(0)$ describes the manner in which the stable and unstable manifolds in the ODE phase space intersect each other at $c = c^*$; i.e., it determines the manner in which the manifolds separate as c is varied from c^*. Thus, the above theorem shows that there is a connection between the stability of the wave and the Melnikov integral $\langle \phi_z, v \rangle$ (see also [30], [32], [33], [34]). If $\lambda_0 = 0$ and g.m. = a.m. $= m \geq 2$, then the PDE (1.2) possesses more than one symmetry, so that $\partial_\lambda^m E(0)$, and hence the Melnikov integrals $\langle \psi_{1,i}, v_j \rangle$, measures the manner in which manifolds separate as the different symmetry parameters are varied. For example, when considering Ginzburg–Landau perturbations of the nonlinear Schrödinger equation, g.m. = a.m. = 2, and the PDE is invariant under both a $SO(2)$ rotation and a spatial translation. For this particular problem, the interested reader should consult Kapitula [15] for a discussion in which the connection is made between $\partial_\lambda^2 E(0)$ and the manner in which the stable and unstable manifolds intersect.

Pego and Weinstein [29] have formulated the Evans function in a manner which is similiar to that presented in Jones [12]. Their formulation is applicable only in the circumstance that the g.m. of an eigenvalue is necessarily one. Physical examples in which g.m. = 1 include the KdV and the KdV–Burgers equations. This formulation is less general than that given by Alexander, Gardner, and Jones [1]. Using Proposition 2.3, it can be shown that Pego and Weinstein's evaluation of $\partial_\lambda E(\lambda_0)$ coincides with that given in Theorem 1.1 in the case that g.m. = a.m. = 1. Pego and Weinstein have also derived expressions for higher derivatives of the Evans function; however, it is not immediately clear that the expressions are equivalent to those presented in Theorem 1.1. Alexander et al. [2] have generalized Pego and Weinstein's approach in the circumstance that g.m. ≥ 2. It is not immediately apparent that one can rederive the results of Theorem 1.1 using the formulation in [2], although one can hypothesize that this is indeed the case.

Now that expressions are available for the derivatives of the Evans function in terms of generalized Melnikov integrals, it will be of interest to determine the manner in which perturbation calculations can be performed. Assuming that some of the eigenvalues move under perturbation, one may naively attempt to take the partial derivative with respect to ϵ of the determinant expression given in Theorem 1.1. This approach is problematic, however, for if g.m. ≥ 2 and $j > m$ the choices of the generalized eigenfunctions must first be known for ϵ nonzero before taking derivatives. Furthermore, if g.m. ≥ 2, it may turn out that generically $\partial_\epsilon \partial_\lambda^j E(\lambda_0) = 0$, but

$\partial_\epsilon^2 \partial_\lambda^j E(\lambda_0) \neq 0$. Using this idea, it might then be necessary to take two derivatives with respect to ϵ of the eigenfunctions. As is seen in section 4 for the case g.m. $= 2$, these technical difficulties can be circumvented.

The linear operator L will be written as

$$L(\epsilon) = B(\epsilon)\partial_z^2 + P(z,\epsilon)\partial_z + N(z,\epsilon),$$

where the $n \times n$ matrices are assumed to be smooth in their arguments. Thus, the perturbation expansions given in section 4 do *not* apply to those problems in which the underlying wave is constructed via geometric singular perturbation theory. It will be assumed that when $\epsilon = 0$ the location and structure of the point eigenvalues for $L(0)$ are completely understood; thus, the zero set of $E(\lambda, 0)$ is assumed to be known. The determination of the eigenvalue structure for $\epsilon \neq 0$ but small, i.e., the zero set of $E(\lambda, \epsilon)$, can then be theoretically accomplished via a Lyapunov–Schmidt reduction. However, if the a.m. of an eigenvalue is larger than its g.m., there could be difficulties, i.e., the eigenvalues may not be smooth functions of ϵ. While this is not problematic if g.m. $= 1$ (see Theorem 1.3), it may cause difficulties if g.m. ≥ 2. As it turns out, g.m. $= 2$ for the nonlinear Schrödinger equation. It is desirable to avoid these potential difficulties if possible. Of course, in doing so some information likely will be lost; however, it may still be possible to recover that lost information via some symmetries present in the eigenvalue problem. This scenario will be illustrated in section 5.1 when the linearized stability of solitary wave solutions to the parametrically forced nonlinear Schrödinger equation is considered.

In section 4 expressions are derived for $\partial_\epsilon^j \partial_\lambda^k E(\lambda_0, 0)$, where λ_0 is the eigenvalue when $\epsilon = 0$ and j and k depend on the a.m. and g.m. of the eigenvalue for both $\epsilon = 0$ and $\epsilon \neq 0$. While the expressions are computed only for g.m. ≤ 2, the algorithm used to generate them can be used to find expressions when g.m. ≥ 3. It is of interest to note the relationship between the various derivatives and the eigenfunction solvability conditions. For example, consider the case that g.m. $= 1$ and a.m. $= p$ when $\epsilon = 0$. In Lemma 4.3 it is shown that

$$\partial_\epsilon E(\lambda_0, 0) = -\langle \partial_\epsilon L(0)\psi_{1,1}, v_1 \rangle.$$

Note that the solvability condition for $\psi_{1,1}$ to remain an eigenfunction for $\epsilon \neq 0$ is

$$\langle \partial_\epsilon L(0)\psi_{1,1}, v_1 \rangle = 0.$$

If $\partial_\epsilon E(\lambda_0, 0) = 0$, so that the eigenvalue persists for $\epsilon \neq 0$, and if a.m. ≥ 2 when $\epsilon = 0$, then as a consequence of Lemma 4.6

$$\partial_\epsilon \partial_\lambda E(\lambda_0, 0) = \langle -\partial_\epsilon L(0)\psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle.$$

It should be noted that setting the right-hand side of the above expression to zero is the solvability condition for $\psi_{2,1}$ to remain an eigenfunction for $\epsilon \neq 0$. Thus, the eigenvalue λ_0 will be simple for $\epsilon \neq 0$ only if the solvability condition for $\psi_{2,1}$ is not satisfied. Examination of the results in section 4.2, where it is assumed that g.m. $= 2$ when $\epsilon = 0$, also shows that there is a strong correlation between the derivatives of the Evans function with respect to ϵ and the eigenfunction solvability conditions.

When g.m. $= 1$, it turns out that the location of all of the eigenvalues can be determined for $\epsilon \neq 0$. This fact is a consequence of the Taylor expansion of the Evans function about $\lambda = \lambda_0$ and a result of Kato [18]. Unfortunately, if g.m. ≥ 2, it is not

clear at the moment as to how one would locate all of these perturbed eigenvalues. This topic will be the focus of a future paper.

THEOREM 1.3. *Suppose that when $\epsilon = 0$, λ_0 is an eigenvalue with g.m. $= 1$ and a.m. $= p$. Further suppose that for $\epsilon \neq 0$, λ_0 is an eigenvalue with a.m. $= j$, where $0 \leq j \leq p-1$. In this case,*

$$\partial_\epsilon \partial_\lambda^j E(\lambda_0, 0) = j! \langle -\partial_\epsilon L(0)\psi_{j+1,1} + \partial_\epsilon \psi_{j,1}, v_1 \rangle,$$

and the location of the remaining $p-j$ eigenvalues is given by

$$\lambda = \lambda_0 + \alpha_1 \omega^h \epsilon^{1/(p-j)} + O(\epsilon^{2/(p-j)}), \quad h = 0, \ldots, p-j-1,$$

where

$$\alpha_1 = -\frac{\langle -\partial_\epsilon L(0)\psi_{j+1,1} + \partial_\epsilon \psi_{j,1}, v_1 \rangle}{\langle \psi_{p,1}, v_1 \rangle}, \quad \omega = e^{2\pi i/(p-j)}.$$

In order to evaluate the expressions given in section 4 for a specific problem, it is necessary to have analytic expressions for the eigenfunctions of $L(0)$. In general, of course, this information will not be available. However, when one is considering the stability of solitary wave solutions to perturbed integrable problems there is a good likelihood that one can indeed compute the eigenfunctions. Examples of such integrable equations include the KdV equation, the focusing and defocusing nonlinear Schrödinger equation, and the Sine–Gordon equation. These eigenfunctions are called the *squared eigenfunctions* [19], [20], [21], [22], [23], [24] and have been used by Pego and Weinstein [29] for calculations with the KdV equation and Kapitula and Sandstede [16], [17] for calculations with perturbed nonlinear Schrödinger equations.

In section 5 the theoretical results of section 4 are applied to perturbed nonlinear Schrödinger equations. When considering the unperturbed problem, g.m. $= 2$ and a.m. $= 4$ for the isolated eigenvalue $\lambda = 0$. Since the nonlinear Schrödinger equation is integrable, analytic expressions are available for the eigenfunctions and adjoint eigenfunctions [36]. For the perturbations of the equation under consideration, the eigenvalue $\lambda = 0$ either becomes simple (section 5.1), ceases to exist (section 5.3), or has g.m. $= 2$ and $2 \leq$ a.m. ≤ 3 for $\epsilon \neq 0$ (section 5.2). It should be noted that in the first case $\partial_\epsilon^2 \partial_\lambda E(0)$ must be calculated, in the second case $\partial_\epsilon^2 E(0)$ must be determined, and in the final case $\partial_\epsilon^2 \partial_\lambda^2 E(0)$ must be found. In each case the first derivative with respect to ϵ is found to be zero.

The paper is organized as follows. In the second section some preliminary results are established. The proof of the theorem is given in the third section. In the fourth section, some perturbation expressions are derived for the Evans function, and in the final section some results are presented for perturbations of the nonlinear Schrödinger equation.

2. Preliminary results. Let $\lambda_0 \in \Omega$ be an eigenvalue of L. Set $T_{\lambda_0} = L - \lambda_0$. The ascent of T_{λ_0} is the smallest number a such that $\mathcal{N}(T_{\lambda_0}^{a+1}) = \mathcal{N}(T_{\lambda_0}^a)$ ($\mathcal{N}$ denotes the null space) (Taylor and Lay [35]). It is known that

$$X = \mathcal{N}(T_{\lambda_0}^a) \oplus \mathcal{R}(T_{\lambda_0}^a),$$

where $X = \mathrm{BU}(\mathbf{R}; \mathbf{R}^n)$ and $\mathcal{R}$ denotes the range space [35].

LEMMA 2.1 (Gardner and Jones [8]). *Let λ_0 be an eigenvalue of g.m. $= m$, a.m. $= p$, and ascent a. Then there exist functions $\psi_{j,i}$, $i = 1, \ldots, m$, $j = 1, \ldots, a_i$,*

with

$$p = \sum_{i=1}^{m} a_i$$

such that

$$T_{\lambda_0}\psi_{j,i} = \psi_{j-1,i}, \quad \psi_{0,i} = 0.$$

Furthermore,

$$\mathcal{N}(T_{\lambda_0}^a) = \text{Span}\{\psi_{j,i}\}, \quad i = 1, \dots, m,\ j = 1, \dots, a_i.$$

An observation yields the following proposition.

PROPOSITION 2.2. *Let $v \in \mathcal{N}(T_{\lambda_0}^*)$. Then*

$$\langle \psi_{j,i}, v\rangle = 0, \quad i = 1, \dots, m,\ j = 1, \dots, a_i - 1.$$

Proof. This follows immediately from

$$\begin{aligned} \langle T_{\lambda_0}\psi_{j,i}, v\rangle &= \langle \psi_{j-1,i}, v\rangle \\ &= \langle \psi_{j,i}, T_{\lambda_0}^* v\rangle \\ &= 0 \end{aligned}$$

as $v \in \mathcal{N}(T_{\lambda_0}^*)$. □

Upon setting $\mathbf{Y} = (u, u')$, where $' = d/dz$, the eigenvalue equation $(L - \lambda)u = 0$ can be rewritten as the first-order system

$$\mathbf{Y}' = M(\lambda, z)\mathbf{Y}, \tag{2.1}$$

where M is the $2n \times 2n$ matrix

$$M(\lambda, z) = \begin{bmatrix} 0 & I_n \\ -B^{-1}(N(z) - \lambda I_n) & -B^{-1}P(z) \end{bmatrix}. \tag{2.2}$$

PROPOSITION 2.3. *Let $\mathbf{Z} = (Z_1, Z_2)^T$ solve the adjoint equation*

$$\mathbf{Z}' = -M^*(\lambda_0, z)\mathbf{Z}.$$

Then

$$T_{\lambda_0}^*(B^{-T}Z_2) = 0,$$

where $B^{-T} = (B^{-1})^T$.

Proof. The operator $T_{\lambda_0}^* = L^* - \lambda_0^*$ is given by

$$(L^* - \lambda_0^*)u = B^T\partial_z^2 u - \partial_z(P^T u) + (N^T - \lambda_0^*)u.$$

An examination of the adjoint equation associated with (2.1), i.e.,

$$\begin{aligned} Z_1' &= (N^T - \lambda_0^* * I_n)B^{-T}Z_2, \\ Z_2' &= -Z_1 + P^T B^{-T} Z_2, \end{aligned}$$

reveals that $\tilde{Z}_2 = B^{-T}Z_2$ does indeed satisfy $(L^* - \lambda_0^*)\tilde{Z}_2 = 0$. □

For the rest of this article,

$$M_\lambda = \partial_\lambda M(\lambda_0, z).$$

LEMMA 2.4. *Let* $\mathbf{Z} = (Z_1, Z_2)^T$ *be a solution to the adjoint equation*

$$\mathbf{Z}' = -M^*(\lambda_0, z)\mathbf{Z}$$

such that $|\mathbf{Z}(z)| \to 0$ *as* $|z| \to \infty$. *Let* $\mathbf{G} : \mathbf{R} \to \mathbf{C}^{2n}$ *be a uniformly bounded continuous function. Then*

$$\langle M_\lambda \mathbf{G}, \mathbf{Z}\rangle = \langle \pi(\mathbf{G}), v\rangle,$$

where $\pi : \mathbf{C}^{2n} \to \mathbf{C}^n$ *is the projection onto the first* n *components and* $v = B^{-T} Z_2 \in \mathcal{N}(T^*_{\lambda_0})$.

Proof. Note that $M_\lambda \mathbf{G} = (0, B^{-1}\pi(\mathbf{G}))^T$. One can then write

$$\begin{aligned} \langle M_\lambda \mathbf{G}, \mathbf{Z}\rangle &= \langle B^{-1}\pi(\mathbf{G}), Z_2\rangle \\ &= \langle \pi(\mathbf{G}), v\rangle, \end{aligned}$$

where $v = B^{-T} Z_2$. As a consequence of Proposition 2.3, the function v is an element of $\mathcal{N}(T^*_{\lambda_0})$. □

Let $\phi_1, \ldots, \phi_{2n}$ be any linearly independent collection of solutions to (2.1). For each i, consider the scaled $(2n-1)$-form

$$e_i = m(\lambda, z)\, \phi_1 \wedge \cdots \wedge \phi_{i-1} \wedge \phi_{i+1} \wedge \cdots \wedge \phi_{2n},$$

where

$$m(\lambda, z) = \exp\left(- \int_0^z \operatorname{tr} M(\lambda, s)\, ds \right).$$

Note that Abel's formula implies that $\phi_i \wedge e_i$ is a nonzero constant for each i. The following was stated in Alexander, Gardner, and Jones [1] and Kapitula [14]; however, the proof will be given here for completeness.

PROPOSITION 2.5. *For each* $i = 1, \ldots, 2n$ *there exists a function* $\mathbf{Y}^{\mathbf{A}}{}_i$ *such that*

$$\mathbf{H} \wedge e_i = \mathbf{H} \cdot \mathbf{Y}^{\mathbf{A}}{}_i,$$

where $\mathbf{H} : \mathbf{R} \to \mathbf{C}^{2n}$ *is a uniformly bounded continuous function. Furthermore, the functions* $\mathbf{Y}^{\mathbf{A}}{}_i$ *form a linearly independent set of solutions to the adjoint equation associated with* (2.1) *and satisfy*

$$\mathbf{Y}^{\mathbf{A}}{}_i \cdot \phi_j = C_i \delta_{ij},$$

where $C_i \neq 0$ *for all* i.

Proof. By the Riesz representation theorem, for each i there exists a function $\mathbf{Y}^{\mathbf{A}}{}_i$ such that

$$\mathbf{H} \wedge e_i = \mathbf{H} \cdot \mathbf{Y}^{\mathbf{A}}{}_i.$$

Since $\phi_j \wedge e_i = C_i \delta_{ij}$, by the above statement

$$\phi_j \cdot \mathbf{Y}^{\mathbf{A}}{}_i = C_i \delta_{ij},$$

where the C_i's are nonzero. Differentiating yields that

$$(\mathbf{Y}^{\mathbf{A}}{}'_i + M^* \mathbf{Y}^{\mathbf{A}}{}_i) \cdot \phi_j = 0,$$

from which one can conclude, since the ϕ_i's are linearly independent, that $\mathbf{Y}^{\mathbf{A}}{}_i$ must be a solution to the adjoint equation. □

3. Proof of Theorem 1.1. For $\lambda \in \Omega$ there exist complex analytic functions $\mathbf{Y}_i(\lambda, z)$, $i = 1, \dots, 2n$, which are solutions to (2.1) and satisfy $|\mathbf{Y}_i(\lambda, z)| \to 0$, $i = 1, \dots, n$, exponentially fast as $z \to -\infty$ and $|\mathbf{Y}_i(\lambda, z)| \to 0$, $i = n+1, \dots, 2n$, exponentially fast as $z \to \infty$. Furthermore, these solutions can be chosen so that the n-forms

$$\mathbf{Y}^u(\lambda, z) = (\mathbf{Y}_1 \wedge \cdots \wedge \mathbf{Y}_n)(\lambda, z), \quad \mathbf{Y}^s(\lambda, z) = (\mathbf{Y}_{n+1} \wedge \cdots \wedge \mathbf{Y}_{2n})(\lambda, z) \tag{3.1}$$

are nonzero for $\lambda \in \Omega$ [1]. The Evans function is then given by

$$E(\lambda) = m(\lambda, z)\, \mathbf{Y}^u(\lambda, z) \wedge \mathbf{Y}^s(\lambda, z), \tag{3.2}$$

where

$$m(\lambda, z) = \exp\left(- \int_0^z \operatorname{tr} M(\lambda, s)\, ds \right).$$

There is a considerable amount of arbitrariness in defining the functions $\mathbf{Y}_i$. Order the functions so that

$$\mathbf{Y}_i(\lambda_0, z) = \mathbf{Y}_{n+i}(\lambda_0, z) = (\psi_{1,i}, \psi_{1,i}')^T, \quad i = 1, \dots, m, \tag{3.3}$$

and assume that this has been done in such a way that the definitions remain consistent for λ near λ_0. The Evans function can then be redefined to be

$$E(\lambda) = m(\lambda, z)\, \mathbf{Y}_1 \wedge \mathbf{Y}_{n+1} \wedge \cdots \wedge \mathbf{Y}_m \wedge \mathbf{Y}_{n+m} \wedge \Phi, \tag{3.4}$$

where $\Phi(\lambda, z) \in \Lambda^{2(n-m)}(\mathbf{C}^{2n}) \neq 0$ for λ near λ_0 is given by

$$\Phi = \mathbf{Y}_{m+1} \wedge \cdots \wedge \mathbf{Y}_n \wedge \mathbf{Y}_{n+m+1} \wedge \cdots \wedge \mathbf{Y}_{2n}.$$

LEMMA 3.1 (Gardner and Jones [8]). *Suppose that the g.m.=m and a.m.=p. Then*

$$\begin{aligned} \partial_\lambda^p E(\lambda_0) &= \frac{p!}{\prod_{i=1}^m a_i!} [\partial_\lambda^{a_1}(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \mathbf{Y}_1] \wedge \cdots \\ &\qquad \wedge [\partial_\lambda^{a_m}(\mathbf{Y}_m - \mathbf{Y}_{n+m}) \wedge \mathbf{Y}_m] \wedge [m(\lambda_0, z)\Phi] \\ &\neq 0, \end{aligned}$$

where

$$p = \sum_{i=1}^m a_i.$$

COROLLARY 3.2 (Gardner and Jones [8]). *Under the assumptions of the above lemma, the functions*

$$\partial_\lambda^{a_i}(\mathbf{Y}_i - \mathbf{Y}_{n+i}), \quad i = 1, \dots, m,$$

grow exponentially fast as $z \to \pm\infty$.

In the subsequent discussion, it will be useful to note that the expression given in Lemma 3.1 can be rewritten as

$$\begin{aligned} \partial_\lambda^p E(\lambda_0) &= (-1)^{m_s} \frac{p!}{\prod_{i=1}^m a_i!}\, \partial_\lambda^{a_1}(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \cdots \wedge \partial_\lambda^{a_m}(\mathbf{Y}_m - \mathbf{Y}_{n+m}) \\ &\qquad \wedge \mathbf{Y}_1 \wedge \cdots \wedge \mathbf{Y}_m \wedge [m(\lambda_0, z)\Phi], \end{aligned} \tag{3.5}$$

where

$$m_s = \sum_{i=0}^{m-1} i.$$

Using Corollary 3.2 as a guide, define the solutions $\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_m$ at $\lambda = \lambda_0$ to be such that the set $\{\tilde{\mathbf{Y}}_1, \ldots, \tilde{\mathbf{Y}}_m\}$ is linearly independent and $|\tilde{\mathbf{Y}}_i(z)| \to +\infty$ as $|z| \to +\infty$. Furthermore, let the solutions be scaled so that when $\lambda = \lambda_0$,

$$D = m(\lambda_0, z)\, \tilde{\mathbf{Y}}_1 \wedge \cdots \wedge \tilde{\mathbf{Y}}_m \wedge \mathbf{Y}_1 \wedge \cdots \wedge \mathbf{Y}_m \wedge \Phi = 1. \tag{3.6}$$

As a consequence of Proposition 2.5 there exist solutions, $\mathbf{Y}^{\mathbf{A}}{}_i$, to the adjoint equation such that

$$\tilde{\mathbf{Y}}_i \cdot \mathbf{Y}^{\mathbf{A}}{}_i = D, \quad i = 1, \ldots, m. \tag{3.7}$$

Since the functions $\tilde{\mathbf{Y}}_i$ grow exponentially fast as $|z| \to +\infty$, by (3.7) these adjoint solutions must satisfy

$$|\mathbf{Y}^{\mathbf{A}}{}_i(z)| \leq C_i e^{-\alpha_i |z|}, \quad i = 1, \ldots, m, \tag{3.8}$$

for some positive constants C_i and α_i. Now define the rest of the adjoint solutions, $\mathbf{Y}^{\mathbf{A}}{}_j$, $j = m+1, \ldots, 2n$, to be such that

$$\begin{aligned} \mathbf{Y}_i \cdot \mathbf{Y}^{\mathbf{A}}{}_{m+i} &= D, \qquad i = 1, \ldots, n, \\ \mathbf{Y}_i \cdot \mathbf{Y}^{\mathbf{A}}{}_i &= D, \qquad i = n+m+1, \ldots, 2n. \end{aligned} \tag{3.9}$$

As a consequence of the fact that the functions $|\mathbf{Y}_i| \to 0$ exponentially fast as $|z| \to \infty$ for $i = 1, \ldots, m$, there exist positive constants C_i and α_i such that

$$2C_i e^{(\alpha_i + \delta)|z|} \geq |\mathbf{Y}^{\mathbf{A}}{}_i(z)| \geq C_i e^{\alpha_i |z|}, \quad i = m+1, \ldots, 2m, \tag{3.10}$$

for some $0 \leq \delta \ll 1$. The remaining adjoint solutions are such that they approach zero exponentially fast in one direction while blowing up exponentially fast in the other.

Let $\mathbf{G}_\pm : \mathbf{R} \to \mathbf{C}^{2n}$ be a uniformly bounded continuous function such that

$$\begin{aligned} |\mathbf{G}_+(z)| &\leq C|z|^k e^{-\alpha z}, \qquad z \geq 1, \\ |\mathbf{G}_-(z)| &\leq C|z|^k e^{\alpha z}, \qquad z \leq -1, \end{aligned} \tag{3.11}$$

for some positive constants C, α, and k. By using variation of parameters, the solution to

$$\mathbf{Y}' = M(\lambda_0, z)\mathbf{Y} + \mathbf{G}_\pm \tag{3.12}$$

is given by

$$\begin{aligned} \mathbf{Y}_\pm = \frac{1}{D} \Bigg(& \sum_{i=1}^{m} c_i^{\pm}(\mathbf{G}_\pm)\tilde{\mathbf{Y}}_i + \sum_{i=1}^{m} \tilde{c}_i(\mathbf{G}_\pm)\mathbf{Y}_i \\ & + \sum_{i=m+1}^{n} c_i(\mathbf{G}_\pm)\mathbf{Y}_i + \sum_{i=n+m+1}^{2n} c_i(\mathbf{G}_\pm)\mathbf{Y}_i \Bigg) \\ & + \sum_{i=1}^{n} d_i^{-}\mathbf{Y}_i + \sum_{i=n+1}^{2n} d_i^{+}\mathbf{Y}_i, \end{aligned} \tag{3.13}$$

where

$$\begin{aligned} c_i^{\pm}(\mathbf{G}_{\pm}) &= \int_{\pm\infty}^{z} \mathbf{G}_{\pm}(s)\cdot \mathbf{Y}^{\mathbf{A}}{}_i(s)\,ds, \quad i=1,\dots,m, \\ \tilde{c}_i(\mathbf{G}_{\pm}) &= \int_{0}^{z} \mathbf{G}_{\pm}(s)\cdot \mathbf{Y}^{\mathbf{A}}{}_{m+i}(s)\,ds, \quad i=1,\dots,m, \\ c_i(\mathbf{G}_{\pm}) &= \begin{cases} \displaystyle\int_{+\infty}^{z} \mathbf{G}_{\pm}(s)\cdot \mathbf{Y}^{\mathbf{A}}{}_{m+i}(s)\,ds, & i=m+1,\dots,n, \\ \displaystyle\int_{-\infty}^{z} \mathbf{G}_{\pm}(s)\cdot \mathbf{Y}^{\mathbf{A}}{}_{i}(s)\,ds, & i=n+m+1,\dots,2n, \end{cases} \end{aligned} \tag{3.14}$$

and $d_i^{\pm}$ are constants [1], [14]. For the solution $\mathbf{Y}_-$ one fixes $d_i^+ = 0$ for $i = n+1,\dots,2n$, while for $\mathbf{Y}_+$ one fixes $d_i^- = 0$ for $i = 1,\dots,n$. As a consequence of the previous estimates and assumption (3.11),

$$\begin{aligned} |\mathbf{Y}_+(z)| &\leq C|z|^{k+1}e^{-\alpha z}, \qquad z \geq 1, \\ |\mathbf{Y}_-(z)| &\leq C|z|^{k+1}e^{\alpha z}, \qquad z \leq -1. \end{aligned} \tag{3.15}$$

LEMMA 3.3. *Define*

$$\mathbf{\Psi}_{j,i} = (\psi_{j,i}, \psi'_{j,i})^T$$

for $i = 1,\dots,m$ and $j = 1,\dots,a_i$. Then for each $i = 1,\dots,m$ and $j = 1,\dots,a_i - 1$,

$$\partial_\lambda^j \mathbf{Y}_i = \partial_\lambda^j \mathbf{Y}_{n+i} = j!\mathbf{\Psi}_{j+1,i}.$$

Proof. See Gardner and Jones [8]. □

Let $1 \leq i \leq m$ be given. For $1 \leq k \leq a_i$, differentiation of (2.1) with respect to λ and evaluating at λ_0 yields

$$(\partial_\lambda^k \mathbf{Y}_\alpha)' = M(\lambda_0,z)\,\partial_\lambda^k \mathbf{Y}_\alpha + kM_\lambda(\lambda_0,z)\,\partial_\lambda^{k-1}\mathbf{Y}_\alpha, \quad \alpha = i, n+i. \tag{3.16}$$

Upon substituting the expression in Lemma 3.3 into (3.16) with $k = a_i$, one then sees that

$$(\partial_\lambda^{a_i} \mathbf{Y}_\alpha)' = M(\lambda_0,z)\,\partial_\lambda^{a_i} \mathbf{Y}_\alpha + M_\lambda(\lambda_0,z)(a_i!\mathbf{\Psi}_{a_i,i}), \quad \alpha = i, n+i. \tag{3.17}$$

Upon requiring that $|\partial_\lambda^{a_i}\mathbf{Y}_i| \to 0$ as $z \to -\infty$ and $|\partial_\lambda^{a_i}\mathbf{Y}_{n+i}| \to 0$ as $z \to +\infty$, and using the solution formula (3.13), after subtracting the solutions one sees that

$$\begin{aligned} \partial_\lambda^{a_i}(\mathbf{Y}_i - \mathbf{Y}_{n+i}) &= \frac{a_i!}{D}\sum_{k=1}^{m} \langle M_\lambda \mathbf{\Psi}_{a_i,i}, \mathbf{Y}^{\mathbf{A}}{}_k\rangle\,\tilde{\mathbf{Y}}_k + \sum_{l=m+1}^{2n} d_l \mathbf{Y}_l \\ &= \frac{a_i!}{D}\sum_{k=1}^{m} \langle \psi_{a_i,i}, v_k\rangle\,\tilde{\mathbf{Y}}_k + \sum_{l=m+1}^{2n} d_l \mathbf{Y}_l, \end{aligned} \tag{3.18}$$

where the d_l's are some constants. In (3.18), $v_k \in \mathcal{N}(T^*_{\lambda_0})$ and the second equality is a consequence of Lemma 2.4.

Now substitute the expressions given in (3.18) into the derivative for the Evans function given in (3.5). Using the fact that D has been normalized to unity, it is then seen that

$$\partial_\lambda^p E(\lambda_0) = (-1)^{m_s} p! \begin{vmatrix} \langle \psi_{a_1,1}, v_1 \rangle & \cdots & \langle \psi_{a_1,1}, v_m \rangle \\ \vdots & & \vdots \\ \langle \psi_{a_m,m}, v_1 \rangle & \cdots & \langle \psi_{a_m,m}, v_m \rangle \end{vmatrix}. \tag{3.19}$$

The proof of Theorem 1.1 is now complete.

4. Perturbation expressions. Now that expressions in terms of the eigenfunctions and adjoint eigenfunctions are available for the derivatives of the Evans function, it will be of interest to determine the manner in which calculations can be performed under perturbation. In this section we will consider only a few special cases, which are motivated by specific problems. However, the ideas presented herein can be generalized. In this section the linear operator L will be written as

$$L(\epsilon) = B(\epsilon)\partial_z^2 + P(z,\epsilon)\partial_z + N(z,\epsilon),$$

where the $n \times n$ matrices are assumed to be smooth in their arguments. Furthermore, it will be assumed that when $\epsilon = 0$, the location and structure of the point eigenvalues for $L(0)$ are completely understood.

Before continuing, the following preliminary lemma will be necessary. In all that follows, the explicit dependence of the matrix M on ϵ will be suppressed. In addition, subscripted variables represent derivatives, and all derivatives will be evaluated at $(\lambda, \epsilon) = (\lambda_0, 0)$.

LEMMA 4.1. *Let $\mathbf{Y} = (\psi, \psi')^T$ be a bounded solution to $\mathbf{Y}' = M(\lambda_0, z)\mathbf{Y}$, and let $\mathbf{Z} = (Z_1, Z_2)^T$ be a bounded solution to the adjoint problem $\mathbf{Z}' = -M(\lambda_0^*, z)\mathbf{Z}$. Then*

$$\langle M_\epsilon \mathbf{Y}, \mathbf{Z} \rangle = -\langle L_\epsilon \psi, v \rangle,$$

where $v = B^{-T} Z_2$ is a bounded solution to the adjoint problem $(L^*(0) - \lambda_0^*)v = 0$. Furthermore,

$$\langle M_{\epsilon\lambda} \mathbf{Y}, \mathbf{Z} \rangle = -\langle B_\epsilon B^{-1} \psi, v \rangle,$$

and if $\mathbf{Y}$ varies smoothly with respect to ϵ, then

$$\langle M_\lambda \mathbf{Y}_\epsilon, \mathbf{Z} \rangle = \langle \partial_\epsilon \psi, v \rangle.$$

Proof. First suppose that $B_\epsilon \neq 0$, so that

$$(B^{-1})_\epsilon = -B^{-1} B_\epsilon B^{-1}.$$

A routine calculation then shows that

$$\langle M_\epsilon \mathbf{Y}, \mathbf{Z} \rangle = \langle B_\epsilon B^{-1}((N - \lambda_0 I_n)\psi + P\psi') - (N_\epsilon \psi + P_\epsilon \psi'), B^{-T} Z_2 \rangle.$$

Since

$$L_\epsilon = B_\epsilon \partial_z^2 + P_\epsilon \partial_z + N_\epsilon$$

and

$$(L(0) - \lambda_0)\psi = B(0)\partial_z^2 \psi + P(z,0)\partial_z \psi + (N(z,0) - \lambda_0 I_n)\psi = 0,$$

upon substitution one sees that

$$B_\epsilon B^{-1}((N-\lambda_0 I_n)\psi + P\psi') - (N_\epsilon \psi + P_\epsilon \psi') = -L_\epsilon \psi,$$

which completes the proof of the first part.

The second part of the lemma follows from the fact that

$$\langle M_{\epsilon\lambda}\mathbf{Y}, \mathbf{Z}\rangle = -\langle B_\epsilon B^{-1}\psi, B^{-T}Z_2\rangle,$$

while the last part follows from

$$\langle M_\lambda \mathbf{Y}_\epsilon, \mathbf{Z}\rangle = \langle \partial_\epsilon \psi, B^{-T}Z_2\rangle.$$

If $B_\epsilon = 0$, then it is routine to check that

$$\begin{aligned}\langle M_\epsilon \mathbf{Y}, \mathbf{Z}\rangle &= \langle -(N_\epsilon\psi + P_\epsilon\psi'), B^{-T}Z_2\rangle \\ &= \langle -L_\epsilon\psi, B^{-T}Z_2\rangle. \qquad \square\end{aligned}$$

The proof of the following corollary is similiar to the one of the above lemma, and is left for the interested reader.

COROLLARY 4.2. *Suppose that when* $\epsilon = 0$

$$(L(0)-\lambda_0)\psi_1 = 0, \quad (L(0)-\lambda_0)\psi_2 = \psi_1.$$

Then for $\mathbf{Y} = (\psi_2, \psi_2')^T$

$$\langle M_\epsilon \mathbf{Y}, \mathbf{Z}\rangle = -\langle L_\epsilon \psi_2, v\rangle + \langle B_\epsilon B^{-1}\psi_1, v\rangle.$$

4.1. G.m. = 1. In this subsection it will be assumed that the eigenvalue λ_0 has g.m. $= 1$ and a.m. $= p$ when $\epsilon = 0$. Thus,

$$(L(0)-\lambda_0)\psi_i = \psi_{i-1}$$

for $i = 1, \dots, p$, with $\psi_0 = 0$. Furthermore, there is a unique (up to scalar multiplication) bounded solution v to $(L^*(0)-\lambda_0^*)v = 0$, and as a consequence of Theorem 1.1

$$\partial_\lambda^p E(\lambda_0) = p!\langle \psi_p, v\rangle.$$

Two situations will be discussed. The first is the case in which λ_0 is not an eigenvalue for $\epsilon \neq 0$. In this scenario it will be desirable to calculate $\partial_\epsilon E(\lambda_0)$. The second case is one in which the eigenvalue becomes simple for nonzero ϵ. It will then be desirable to know $\partial_\epsilon \partial_\lambda E(\lambda_0)$.

LEMMA 4.3. *Suppose that* λ_0 *is not an eigenvalue for* $\epsilon \neq 0$. *Then*

$$\partial_\epsilon E(\lambda_0) = -\langle L_\epsilon \psi_1, v\rangle.$$

Proof. First recall from (3.5) that

$$\partial_\lambda^p E(\lambda_0) = \partial_\lambda^p(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \mathbf{Y}_1 \wedge [m(\lambda_0, z)\Phi].$$

In a similiar fashion it is not difficult to show that

$$\partial_\epsilon E(\lambda_0) = \partial_\epsilon(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \mathbf{Y}_1 \wedge [m(\lambda_0, z)\Phi].$$

The equation for $\partial_\epsilon \mathbf{Y}_\alpha$, $\alpha \in \{1, n+1\}$, is given by

$$(\partial_\epsilon \mathbf{Y}_\alpha)' = M(\lambda_0, z)\partial_\epsilon \mathbf{Y}_\alpha + M_\epsilon(\lambda_0, z)\mathbf{Y}_\alpha.$$

Upon requiring that $|\partial_\epsilon \mathbf{Y}_1| \to 0$ as $z \to -\infty$ and $|\partial_\epsilon \mathbf{Y}_{n+1}| \to 0$ as $z \to +\infty$, and by using the ideas leading to (3.13) and (3.18), it can be seen that

$$\partial_\epsilon(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = \frac{1}{D}\left(\langle M_\epsilon \mathbf{Y}_1, \mathbf{Y^A}_1\rangle \tilde{\mathbf{Y}}_1 + \sum_{i=2}^{2n} c_i \mathbf{Y}_i\right).$$

Thus, upon substituting this expression into that for $\partial_\epsilon E(\lambda_0)$ and using the fact that $D = 1$, one gets that

$$\partial_\epsilon E(\lambda_0) = \langle M_\epsilon \mathbf{Y}_1, \mathbf{Y^A}_1\rangle,$$

which, as a consequence of Lemma 4.1, yields the final result. □

Remark 4.4. If λ_0 is a simple eigenvalue, then as a consequence of Theorem 1.1, the above lemma, and the implicit function theorem, it is seen that

$$\partial_\epsilon \lambda_0 = \frac{\langle L_\epsilon \psi_1, v\rangle}{\langle \psi_1, v\rangle}.$$

Remark 4.5. Using standard perturbation theory, ψ_1 will be an eigenfunction of $L(\epsilon)$ with corresponding eigenvalue λ_0 if $\partial_\epsilon E(\lambda_0) = 0$.

LEMMA 4.6. *Suppose that $p \geq 2$ and that $E(\lambda_0) = 0$ for $\epsilon \neq 0$. Then*

$$\partial_\epsilon \partial_\lambda E(\lambda_0) = \langle -L_\epsilon \psi_2 + \partial_\epsilon \psi_1, v\rangle.$$

Proof. Since $p \geq 2$, it is a consequence of Lemma 3.3 that

$$\partial_\lambda(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = 0;$$

thus, a routine calculation shows that

$$\partial_\epsilon \partial_\lambda E(\lambda_0) = \partial^2_{\epsilon\lambda}(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \mathbf{Y}_1 \wedge [m(\lambda_0, z)\Phi].$$

Now,

$$(\partial^2_{\epsilon\lambda}\mathbf{Y}_\alpha)' = M\partial^2_{\epsilon\lambda}\mathbf{Y}_\alpha + M_{\epsilon\lambda}\mathbf{Y}_\alpha + M_\epsilon \partial_\lambda \mathbf{Y}_\alpha + M_\lambda \partial_\epsilon \mathbf{Y}_\alpha$$

for $\alpha \in \{1, n+1\}$. Upon requiring that $\partial^2_{\epsilon\lambda}\mathbf{Y}_1$ decay as $z \to -\infty$ and $\partial^2_{\epsilon\lambda}\mathbf{Y}_{n+1}$ decay as $z \to +\infty$, one then sees that

$$\partial^2_{\epsilon\lambda}(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = \frac{1}{D}\left((c_1^- - c_1^+)\tilde{\mathbf{Y}}_1 + \sum_{i=2}^{2n} d_i \mathbf{Y}_i\right), \tag{4.1}$$

where

$$c_1^- - c_1^+ = \langle M_{\epsilon\lambda}\mathbf{Y}_1 + M_\epsilon \mathbf{\Psi}_2 + M_\lambda \partial_\epsilon \mathbf{Y}_1, \mathbf{Y^A}_1\rangle.$$

In the above expression, recall that Lemma 3.3 states that $\partial_\lambda \mathbf{Y}_1 = \partial_\lambda \mathbf{Y}_{n+1} = \mathbf{\Psi}_2$. Upon substituting back into the expression for $\partial_\epsilon \partial_\lambda E(\lambda_0)$ and using the fact that $D = 1$, it is now seen that

$$\partial_\epsilon \partial_\lambda E(\lambda_0) = c_1^- - c_1^+.$$

Upon using the results of Lemma 4.1 and Corollary 4.2, one gets

$$c_1^- - c_1^+ = -\langle B_\epsilon B^{-1}\psi_1, v\rangle - \langle L_\epsilon \psi_2, v\rangle + \langle B_\epsilon B^{-1}\psi_1, v\rangle + \langle \partial_\epsilon \psi_1, v\rangle.$$

The proof is now complete. □

Remark 4.7. Using standard perturbation theory, if $\partial_\epsilon \partial_\lambda E(\lambda_0) = 0$, then ψ_2 will remain as a generalized eigenfunction of $L(\epsilon)$ for $\epsilon \neq 0$.

By following the proofs of the above lemmas, the perturbation results can clearly be generalized for the case that a.m. $= j < p$ for $\epsilon \neq 0$. Furthermore, since g.m. $= 1$, the location of the remaining eigenvalues can also be given. The proof of Theorem 1.3 is now complete.

THEOREM 4.8. *Suppose that for $\epsilon \neq 0$ λ_0 is an eigenvalue with a.m. $= j$, where $0 \leq j \leq p-1$. In this case,*

$$\partial_\epsilon \partial_\lambda^j E(\lambda_0) = j!\langle -L_\epsilon \psi_{j+1} + \partial_\epsilon \psi_j, v\rangle.$$

Furthermore, the location of the remaining $p-j$ eigenvalues is given by

$$\lambda = \lambda_0 + \alpha_1 \omega^h \epsilon^{1/(p-j)} + O(\epsilon^{2/(p-j)}), \quad h = 0, \ldots, p-j-1,$$

where

$$\alpha_1 = -\frac{\langle -L_\epsilon \psi_{j+1} + \partial_\epsilon \psi_j, v\rangle}{\langle \psi_p, v\rangle}, \quad \omega = e^{2\pi i/(p-j)}.$$

Proof. The perturbation expression is a generalization of that given in the above lemmas, and hence the proof will be left to the interested reader. In order to prove the eigenvalue perturbation expression, consider the following argument. Set $\gamma = \lambda - \lambda_0$. For γ and ϵ small, the Evans function has the Taylor expansion

$$E(\lambda) = \gamma^j \left(\langle -L_\epsilon \psi_{j+1} + \partial_\epsilon \psi_j, v\rangle \epsilon + \cdots + \langle \psi_p, v\rangle \gamma^{p-j} + O(|\gamma|^{p-j+1}) \right),$$

where all the coefficients below γ^{p-j} are $O(\epsilon)$. Following Kato [18], the remaining zeros of the Evans function will be $O(\epsilon^{1/(p-j)})$. As such, it is then easy to see that they must be given by

$$\gamma = \alpha_1 \omega^h \epsilon^{1/(p-j)} + O(\epsilon^{2/(p-j)}), \quad h = 0, \ldots, p-j-1,$$

where

$$\alpha_1 = -\frac{\langle -L_\epsilon \psi_{j+1} + \partial_\epsilon \psi_j, v\rangle}{\langle \psi_p, v\rangle}, \quad \omega = e^{2\pi i/(p-j)}.$$

The result is now proved. □

4.2. G.m. = 2. In this subsection it will be assumed that when $\epsilon = 0$ the eigenvalue λ_0 has g.m. $= 2$ and a.m. $= p \geq 2$. Thus, we have two chains:

$$(L(0) - \lambda_0)\psi_{j,i} = \psi_{j-1,i}, \quad \psi_{0,i} = 0$$

for $i = 1, 2$ and $j = 1, \ldots, a_i$, with $a_1 + a_2 = p$. Without loss of generality it will be assumed in this section that $1 \leq a_1 \leq a_2$. As a consequence of Theorem 1.1, there exists two bounded adjoint solutions, v_1 and v_2, such that

$$\partial_\lambda^p E(\lambda_0) = -p! \begin{vmatrix} \langle \psi_{a_1,1}, v_1\rangle & \langle \psi_{a_1,1}, v_2\rangle \\ \langle \psi_{a_2,2}, v_1\rangle & \langle \psi_{a_2,2}, v_2\rangle \end{vmatrix}.$$

LEMMA 4.9. *Suppose that λ_0 is not an eigenvalue of $L(\epsilon)$ for $\epsilon \neq 0$. Then*

$$\partial_\epsilon E(\lambda_0) = 0,$$

and

$$\partial_\epsilon^2 E(\lambda_0) = -2 \begin{vmatrix} \langle L_\epsilon \psi_{1,1}, v_1 \rangle & \langle L_\epsilon \psi_{1,1}, v_2 \rangle \\ \langle L_\epsilon \psi_{1,2}, v_1 \rangle & \langle L_\epsilon \psi_{1,2}, v_2 \rangle \end{vmatrix}.$$

Proof. It is known that $\mathbf{Y}_1 = \mathbf{Y}_{n+1} = (\psi_{1,1}, \psi'_{1,1})^T$ and that $\mathbf{Y}_2 = \mathbf{Y}_{n+2} = (\psi_{1,2}, \psi'_{1,2})^T$. Using this information, a simple calculation shows that $\partial_\epsilon E(\lambda_0) = 0$.

A standard calculation shows that

$$\partial_\epsilon^2 E(\lambda_0) = -2\partial_\epsilon(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \partial_\epsilon(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) \wedge \mathbf{Y}_1 \wedge \mathbf{Y}_2 \wedge [m(\lambda_0, z)\Phi].$$

Since

$$(\partial_\epsilon \mathbf{Y})' = M \partial_\epsilon \mathbf{Y} + M_\epsilon \mathbf{Y},$$

by following the ideas presented in the previous section it can be seen that

$$\partial_\epsilon(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = \frac{1}{D}\left(\langle -L_\epsilon \psi_{1,1}, v_1 \rangle \tilde{\mathbf{Y}}_1 + \langle -L_\epsilon \psi_{1,1}, v_2 \rangle \tilde{\mathbf{Y}}_2 + \sum_{i=3}^{2n} c_i \mathbf{Y}_i \right)$$

and that

$$\partial_\epsilon(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) = \frac{1}{D}\left(\langle -L_\epsilon \psi_{1,2}, v_1 \rangle \tilde{\mathbf{Y}}_1 + \langle -L_\epsilon \psi_{1,2}, v_2 \rangle \tilde{\mathbf{Y}}_2 + \sum_{i=3}^{2n} d_i \mathbf{Y}_i \right). \tag{4.2}$$

Upon substitution into the expression for $\partial_\epsilon^2 E(\lambda_0)$, and using the fact that $D = 1$, the conclusion of the lemma follows. □

Remark 4.10. The straightforward generalization of Lemma 4.9 to the case when g.m. ≥ 3 will be left to the interested reader.

For the rest of this section the primary interest will be the case that $a_1 > 1$. This is due to the fact that for the applications of interest in this paper, i.e., perturbations of the nonlinear Schrödinger equation, it turns out that $a_1 = a_2 = 2$. If $a_1 = 1$, the final results will be stated; however, the proof will be left to the interested reader.

LEMMA 4.11. *Suppose that for $\epsilon \neq 0$ λ_0 is a simple eigenvalue of $L(\epsilon)$. Further suppose that the eigenfunction for $\epsilon \neq 0$ is $\psi_{1,1}$. Then if $a_1 > 1$,*

$$\partial_\epsilon \partial_\lambda E(\lambda_0) = 0,$$

and

$$\partial_\epsilon^2 \partial_\lambda E(\lambda_0) = 2 \begin{vmatrix} \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle & \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_2 \rangle \\ \langle L_\epsilon \psi_{1,2}, v_1 \rangle & \langle L_\epsilon \psi_{1,2}, v_2 \rangle \end{vmatrix}.$$

Remark 4.12. If $a_1 = 1$, then substitution of the expression for $\partial_\epsilon(\mathbf{Y}_2 - \mathbf{Y}_{n+2})$ into (4.3) yields that

$$\partial_\epsilon \partial_\lambda E(\lambda_0) = \begin{vmatrix} \langle \psi_{1,1}, v_1 \rangle & \langle \psi_{1,1}, v_2 \rangle \\ \langle L_\epsilon \psi_{1,2}, v_1 \rangle & \langle L_\epsilon \psi_{1,2}, v_2 \rangle \end{vmatrix}.$$

Proof. The proof will be only sketched, as it has many similiar features to those given above.

Since $\psi_{1,1}$ is an eigenfunction for $\epsilon \neq 0$, a routine calculation shows that

$$\langle L_\epsilon \psi_{1,1}, v_1 \rangle = \langle L_\epsilon \psi_{1,1}, v_2 \rangle = 0.$$

This is the solvability condition needed in order to solve the equation $\partial_\epsilon \psi_{1,1} = -(L - \lambda_0)^{-1} L_\epsilon \psi_{1,1}$.

First note that

$$\partial_\epsilon \partial_\lambda E(\lambda_0) = -\partial_\lambda(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \partial_\epsilon(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) \wedge \mathbf{Y}_1 \wedge \mathbf{Y}_2 \wedge [m(\lambda_0, z)\Phi]. \tag{4.3}$$

Since $a_1 > 1$, by Lemma 3.3

$$\partial_\lambda(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = 0,$$

from which one immediately sees that $\partial_\epsilon \partial_\lambda E(\lambda_0) = 0$. Taking another derivative with respect to ϵ, it can be seen that

$$\partial_\epsilon^2 \partial_\lambda E(\lambda_0) = -2\, \partial_{\epsilon\lambda}(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \partial_\epsilon(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) \wedge \mathbf{Y}_1 \wedge \mathbf{Y}_2 \wedge [m(\lambda_0, z)\Phi].$$

A slight modification of the argument leading to (4.1) yields that

$$\begin{aligned} &\partial_{\epsilon\lambda}^2(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \\ &\quad = \frac{1}{D}\left(\langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle \tilde{\mathbf{Y}}_1 + \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_2 \rangle \tilde{\mathbf{Y}}_2 + \sum_{i=3}^{2n} d_i \mathbf{Y}_i \right). \end{aligned}$$

After recalling the expression for $\partial_\epsilon(\mathbf{Y}_2 - \mathbf{Y}_{n+2})$ given in (4.2), and using the fact that $D = 1$, upon substitution into the above expression for $\partial_\epsilon^2 \partial_\lambda E(\lambda_0)$ the final result is seen. □

LEMMA 4.13. *Suppose that for $\epsilon \neq 0$ λ_0 is an eigenvalue of $L(\epsilon)$ with g.m. = a.m. = 2. Then if $a_1 > 1$,*

$$\partial_\epsilon \partial_\lambda^2 E(\lambda_0) = 0,$$

and

$$\partial_\epsilon^2 \partial_\lambda^2 E(\lambda_0) = -4 \begin{vmatrix} \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle & \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_2 \rangle \\ \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_1 \rangle & \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_2 \rangle \end{vmatrix}.$$

Remark 4.14. If $a_1 = 1$, then the above expression reduces to

$$\partial_\epsilon \partial_\lambda^2 E(\lambda_0) = -2 \begin{vmatrix} \langle \psi_{1,1}, v_1 \rangle & \langle \psi_{1,1}, v_2 \rangle \\ \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_1 \rangle & \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_2 \rangle \end{vmatrix}.$$

Proof. This proof will be even sketchier than the previous one, since the idea is essentially the same. Since $a_2 \geq a_1 > 1$, by Lemma 3.3 it is known that

$$\partial_\lambda(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = \partial_\lambda(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) = 0.$$

Using this information, it can eventually be concluded that $\partial_\epsilon \partial_\lambda^2 E(\lambda_0) = 0$ and that

$$\partial_\epsilon^2 \partial_\lambda^2 E(\lambda_0) = -4\, \partial_{\epsilon\lambda}^2(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \partial_{\epsilon\lambda}^2(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) \wedge \mathbf{Y}_1 \wedge \mathbf{Y}_2 \wedge [m(\lambda_0, z)\Phi].$$

The expression for $\partial^2_{\epsilon\lambda}(\mathbf{Y}_1 - \mathbf{Y}_{n+1})$ is exactly that as given in the previous proof. The expression for $\partial^2_{\epsilon\lambda}(\mathbf{Y}_2 - \mathbf{Y}_{n+2})$ is found simply by substituting $\psi_{1,2}$ for $\psi_{1,1}$ and $\psi_{2,2}$ for $\psi_{2,1}$ in that expression. Using the definition of $\partial^p_\lambda E(\lambda_0)$ then yields the final result. □

Clearly, this process can be continued ad nauseam, in that one can derive expressions for the case that g.m. = 3 or higher and the various subcases. This process will be left for the interested reader. This section will close with one final lemma, the result of which is needed for one of the applications considered in the following section. For the following lemma, note that if $a_2 = 2$ for all ϵ, so that $(L(\epsilon) - \lambda_0)\psi_{1,2} = 0$ and $(L(\epsilon) - \lambda_0)\psi_{2,2} = \psi_{1,2}$, then it is necessarily true that

$$\langle L_\epsilon \psi_{1,2}, v\rangle = \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v\rangle = 0 \tag{4.4}$$

for all bounded adjoint solutions v. Note that this implies that $\partial^2_\epsilon \partial^2_\lambda E(\lambda_0) = 0$.

LEMMA 4.15. *Suppose that $a_2 = 2$ for all $\epsilon \geq 0$, and that g.m. $= 2$ for $\epsilon \neq 0$. Further suppose that $a_1 > 1$ when $\epsilon = 0$. Then*

$$\partial_\epsilon \partial^3_\lambda E(\lambda_0) = -6 \begin{vmatrix} \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1\rangle & \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_2\rangle \\ \langle \psi_{2,2}, v_1\rangle & \langle \psi_{2,2}, v_2\rangle \end{vmatrix}.$$

Proof. Again, it is only sketched. Since $a_1 > 1$, it is necessarily true that $\partial_\lambda(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) = 0$. This then implies that

$$\partial_\epsilon \partial^3_\lambda E(\lambda_0) = -3\, \partial^2_{\epsilon\lambda}(\mathbf{Y}_1 - \mathbf{Y}_{n+1}) \wedge \partial^2_\lambda(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) \wedge \mathbf{Y}_1 \wedge \mathbf{Y}_2 \wedge [m(\lambda_0, z)\Phi].$$

Since $a_2 = 2$ for all ϵ, $\partial^2_\lambda(\mathbf{Y}_2 - \mathbf{Y}_{n+2}) \neq 0$ is given by (3.18), so that the above expression will be generically nonzero. Substituting in the expression for $\partial^2_{\epsilon\lambda}(\mathbf{Y}_1 - \mathbf{Y}_{n+1})$ yields the final result. □

5. Examples. In this section the theory of the previous sections will be applied to various perturbations of the nonlinear Schrödinger equation (NLS),

$$i\partial_t \phi + (\partial_x^2 - \omega)\phi + 4|\phi|^2\phi = \epsilon R(x, \phi, \phi^*), \tag{5.1}$$

where the perturbation term will be assumed to be smooth. When $\epsilon = 0$, there is a real-valued solitary wave solution given by

$$\Phi(x, \omega) = \sqrt{\frac{\omega}{2}}\, \mathrm{sech}(\sqrt{\omega}\, x). \tag{5.2}$$

Breaking ϕ into its real and imaginary parts and linearizing about the wave, one gets the linear operator L given by

$$L = \begin{bmatrix} 0 & -L_i \\ L_r & 0 \end{bmatrix}, \tag{5.3}$$

where

$$L_i = \partial_x^2 - \omega + 4\Phi^2, \quad L_r = \partial_x^2 - \omega + 12\Phi^2.$$

Note that $L_i(\Phi) = 0$ and $L_r(\partial_x \Phi) = 0$. Following Weinstein [36], one has the following information concerning the operator L.

PROPOSITION 5.1. *The eigenfunctions for the operator L are given by*

$$\psi_{1,1} = \begin{bmatrix} \partial_x \Phi \\ 0 \end{bmatrix}, \quad \psi_{2,1} = \begin{bmatrix} 0 \\ -\frac{1}{2} x \Phi \end{bmatrix}$$

and

$$\psi_{1,2} = \begin{bmatrix} 0 \\ \Phi \end{bmatrix}, \quad \psi_{2,2} = \begin{bmatrix} \partial_\omega \Phi \\ 0 \end{bmatrix}.$$

Furthermore, the bounded solutions to the adjoint equation are given by

$$v_1 = \frac{4}{\langle \Phi, \Phi \rangle} \begin{bmatrix} 0 \\ \partial_x \Phi \end{bmatrix}, \quad v_2 = \frac{2}{\partial_\omega \langle \Phi, \Phi \rangle} \begin{bmatrix} \Phi \\ 0 \end{bmatrix}.$$

Note that v_1 and v_2 have been scaled so that

$$\langle \psi_{2,1}, v_1 \rangle = \langle \psi_{2,2}, v_2 \rangle = 1.$$

Upon using the results of Theorem 1.1 one is able to conclude the following information about the Evans function at $\lambda = 0$. The information for $\lambda \neq 0$ follows immediately from the fact that the wave Φ is a stable solution to (5.1) [37].

LEMMA 5.2. *The Evans function for the operator L satisfies*

$$\partial_\lambda^i E(0) = 0, \quad i = 0, \dots, 3,$$

and

$$\partial_\lambda^4 E(0) = -24.$$

Furthermore, $E(\lambda) < 0$ for $Re\, \lambda > 0$.

For $\epsilon \neq 0$ the wave (assuming that it persists) will in general be complex-valued. After breaking it into its real and complex parts, it will henceforth be denoted by $\mathbf{\Phi}(x, \omega, \epsilon)$, with $\mathbf{\Phi}(x, \omega, 0) = (\Phi, 0)^T$.

5.1. Parametrically forced NLS. For the parametrically forced NLS (PFNLS) the perturbation term is given by

$$R(x, \phi, \phi^*) = -i(\gamma \phi - \mu \phi^* e^{-i2\theta}),$$

where $\gamma, \mu > 0$ and

$$\cos 2\theta = \frac{\gamma}{\mu}.$$

Here γ is the dissipation factor (linear loss) and μ is the parametric gain. For $\epsilon > 0$ the wave is still real-valued and is given by $\mathbf{\Phi} = (\Phi, 0)^T$, where

$$\Phi(x, \omega, \epsilon) = \sqrt{\frac{\beta}{2}}\, \mathrm{sech}(\sqrt{\beta}\, x), \tag{5.4}$$

and

$$\beta = \omega + \epsilon \mu \sin 2\theta. \tag{5.5}$$

Due to the fact that the perturbation R breaks the rotation symmetry, when $\epsilon > 0$ it will generically be true that $\partial_\lambda E(0) \neq 0$. Furthermore, the fact that the translation symmetry is not broken implies that $\psi_{1,1}$ remains as an eigenfunction for L. Since $a_1 = a_2 = 2$ at $\lambda = 0$ when $\epsilon = 0$, the result of Lemma 4.11 will then apply.

After linearizing the PFNLS about the wave, a simple calculation shows that

$$L_\epsilon = \begin{bmatrix} 0 & -\mu \sin 2\theta \\ -\mu \sin 2\theta & -2\gamma \end{bmatrix} + 4\partial_\epsilon(\Phi^2) \begin{bmatrix} 0 & -1 \\ 3 & 0 \end{bmatrix}.$$

Since

$$\partial_\epsilon(\Phi^2) = \frac{\mu \sin 2\theta}{\omega} \left(\Phi^2 + \frac{1}{2} x \partial_x(\Phi^2) \right)$$

and Φ is even in x, it is a routine calculation to see that

$$\langle \partial_\epsilon \psi_{1,1}, v_1 \rangle = \langle L_\epsilon \psi_{1,2}, v_1 \rangle = \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_2 \rangle = 0.$$

Thus, as a consequence of Lemma 4.11

$$\partial_\epsilon^2 \partial_\lambda E(0) = -2 \langle L_\epsilon \psi_{2,1}, v_1 \rangle \langle L_\epsilon \psi_{1,2}, v_2 \rangle. \tag{5.6}$$

Since

$$\langle x\Phi^{2k}, \partial_x(\Phi^2) \rangle = -\frac{1}{k+1} \langle \Phi^{2k}, \Phi^2 \rangle, \quad \langle \Phi, \Phi \rangle = \omega^{1/2}, \quad \langle \Phi^2, \Phi^2 \rangle = \frac{1}{3} \omega^{3/2},$$

one can see that

$$\langle L_\epsilon \psi_{1,2}, v_2 \rangle = -8\omega\mu \sin 2\theta.$$

In addition, the fact that $\langle x\Phi, \partial_x \Phi \rangle = -1/2 \langle \Phi, \Phi \rangle$ implies that

$$\langle L_\epsilon \psi_{2,1}, v_1 \rangle = -2\gamma.$$

Therefore, by (5.6)

$$\partial_\epsilon^2 \partial_\lambda E(0) = -32\omega\gamma\mu \sin 2\theta. \tag{5.7}$$

LEMMA 5.3. *Consider the PFNLS. Suppose that* $0 < \epsilon \ll 1$. *If* $\sin 2\theta < 0$, *then the wave* Φ *is unstable. If* $\sin 2\theta > 0$, *then there are no positive eigenvalues which are of* $O(\epsilon)$.

Proof. The eigenvalues of L are symmetric about the lines $\operatorname{Re} \lambda = -\epsilon\gamma$ and $\operatorname{Im} \lambda = 0$. This is easily seen by noticing that $(L - \lambda)P = 0$ reduces to $L_i L_r P_1 = -\lambda(\lambda + 2\epsilon\gamma)P_1$ and $L_r L_i P_2 = -\lambda(\lambda + 2\epsilon\gamma)P_2$, where $P = (P_1, P_2)^T$.

If $\sin 2\theta < 0$, then as a consequence of (5.7), $\partial_\lambda E(0) > 0$ for $\epsilon > 0$ sufficiently small. By Lemma 5.2, it is necessarily true that $E(\lambda)$ is negative for real λ sufficiently large. This necessarily implies the existence of a real positive zero of $E(\lambda)$, which in turn implies the existence of a real positive eigenvalue.

Now suppose that $\sin 2\theta > 0$, so that $\partial_\lambda E(0) < 0$ for $\epsilon > 0$ sufficiently small. If there is a zero with positive real part, then there must be at two such zeros. Suppose that two such zeros exist. Then there must also exist two other zeros with negative real part which are symmetric with respect to the line $\operatorname{Re} \lambda = -\epsilon\gamma$. Thus, recalling that $\lambda = 0$ is always a zero, there must exist at least five zeros which are of $O(\epsilon)$. This contradicts the fact that there can only be four such zeros. ∎

Remark 5.4. Barashenkov, Bogdan, and Korobov [5] have previously shown that $\sin 2\theta < 0$ implies the existence of an unstable eigenvalue. Their proof, however, is a consequence of the results present in either Grillakis [10] or in Jones [13].

Remark 5.5. By (5.7) it is known that $\partial_\lambda E(0) > 0$ for $\epsilon > 0$ sufficiently small. It can be shown, however, that $\partial_\lambda E(0) > 0$ for all $\epsilon > 0$ [2], [17], [25].

5.2. Cubic-quintic NLS. For the cubic-quintic NLS (CQNLS) the perturbation term is given by

$$R(x,\phi,\phi^*) = i(d_1\partial_x^2\phi + d_2\phi + d_3|\phi|^2\phi + d_4|\phi|^4\phi).$$

The positive parameter d_1 describes spectral filtering, d_2 describes the linear gain ($d_2 > 0$) or loss ($d_2 < 0$) due to the fiber, and d_3 and d_4 describe the nonlinear gain or loss due to the fiber. Note that this perturbation preserves both the rotation and translation symmetry. One then generically expects that $\partial_\lambda^2 E(0) \neq 0$, so that the results of Lemma 4.13, and possibly Lemma 4.15, will apply.

In order for the solution $\phi = 0$ to be stable, one must require that $d_2 < 0$. Since this is a minimal requirement for the wave to be stable, this assumption will hold for the rest of the discussion. It is shown in Kapitula [15] that the wave persists if

$$H(x) = d_1\partial_x^2\Phi + d_2\Phi + d_3\Phi^3 + d_4\Phi^5 \tag{5.8}$$

satisfies

$$\langle H, \Phi\rangle = 0, \tag{5.9}$$

i.e.,

$$d_1 - \frac{3}{\omega}d_2 - d_3 - \frac{2}{5}\omega d_4 = 0.$$

This condition ensures that $L_i^{-1}(H)$ is a bounded function for all x. Using standard perturbation theory, and the fact that

$$L^{-1} = \begin{bmatrix} 0 & L_r^{-1} \\ -L_i^{-1} & 0 \end{bmatrix},$$

it is not difficult to see that

$$\partial_\epsilon \mathbf{\Phi} = \begin{bmatrix} 0 \\ L_i^{-1}(H) \end{bmatrix}. \tag{5.10}$$

The two solutions to $L_i w = 0$ are given by

$$w_1 = \Phi, \quad w_2 = x\Phi + \frac{1}{\sqrt{2}}\sinh(\sqrt{\omega}\,x);$$

thus, upon using variation of parameters one can explicitly construct $L^{-1}(H)$. This will not be done here, however, and will be left as an exercise for the interested reader. It is enough to recognize that $\partial_\epsilon\Phi$ is an even function in x.

Upon linearizing the CQNLS about the wave, a simple calculation shows that

$$L_\epsilon = (d_1\partial_x^2 + d_2)I_2 + d_3\Phi^2\begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix} + d_4\Phi^4\begin{bmatrix} 5 & 0 \\ 0 & 1 \end{bmatrix} + 8\Phi L_i^{-1}(H)\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Furthermore, another calculation shows that

$$\partial_\epsilon\psi_{1,1} = \begin{bmatrix} 0 \\ \partial_x L_i^{-1}(H) \end{bmatrix}, \quad \partial_\epsilon\psi_{1,2} = \begin{bmatrix} -L_i^{-1}(H) \\ 0 \end{bmatrix}.$$

Using the fact that Φ and $L_i^{-1}(H)$ are even in x, it is a trivial matter to show that

$$\langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_2 \rangle = \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_1 \rangle = 0.$$

Therefore, after using the result of Lemma 4.13 it can be seen that

$$\partial_\epsilon^2 \partial_\lambda^2 E(0) = -4 \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle \langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_2 \rangle.$$

The above expressions are difficult to compute, and as such the problem was considered in Kapitula [15] for a more general perturbation of the NLS. In the following, let C_i denote a positive constant. Assuming relation (5.9), i.e., after solving for d_3 in terms of the other variables, as a consequence of the calculations in [15] it can be shown that

$$\langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle = C_1 d_1$$

and that

$$\langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_2 \rangle = C_2 \left(d_2 - \frac{2}{15} \omega^2 d_4 \right).$$

As a consequence of Lemma 4.15, when

$$\langle -L_\epsilon \psi_{2,2} + \partial_\epsilon \psi_{1,2}, v_2 \rangle = 0,$$

then

$$\begin{aligned} \partial_\epsilon \partial_\lambda^3 E(0) &= -6 \langle -L_\epsilon \psi_{2,1} + \partial_\epsilon \psi_{1,1}, v_1 \rangle \\ &= -6 C_1 \, d_1. \end{aligned}$$

It is now possible to describe the location of all of the $O(\epsilon)$ eigenvalues.

LEMMA 5.6. *Let d_3 be such that the wave exists for $0 < \epsilon \ll 1$. Furthermore, suppose that $d_1 > 0$ and $d_2 < 0$. Then if*

$$d_2 < \frac{2}{15} \omega^2 d_4,$$

there exists a real positive eigenvalue and a real negative eigenvalue, both of which are $O(\epsilon)$. If

$$\frac{2}{15} \omega^2 d_4 < d_2 < 0,$$

then both $O(\epsilon)$ eigenvalues are real and negative.

Proof. From the above arguments it is seen that

$$\partial_\epsilon^2 \partial_\lambda^2 E(0) = -4 C_1 C_2 \, d_1 \left(d_2 - \frac{2}{15} \omega^2 d_4 \right),$$

and if $d_2 - 2/15 \, \omega^2 d_4 = 0$, then

$$\partial_\epsilon \partial_\lambda^3 E(0) = -6 C_1 \, d_1.$$

All that is left to recognize is that if $\partial_\lambda^2 E(0) > 0$, then since $\partial_\lambda^4 E(0) < 0$ there exists one positive real zero of $E(\lambda)$ which is of $O(\epsilon)$. When $\partial_\lambda^2 E(0) = 0$, the fact that $\partial_\lambda^3 E(0) < 0$ implies that the zero is moving to the left. ∎

Remark 5.7. For a more complete discussion, the interested reader should consult Kapitula [15] and Kapitula and Sandstede [17].

5.3. Nonhomogeneous PFNLS. As a final example, consider the nonhomogeneous PFNLS (NPFNLS), where the perturbation term is given by

$$R(x,\phi,\phi^*) = -i(\gamma\phi - \mu h(x)\phi^* e^{-i2\theta}).$$

Here $h(x)$ is assumed to be even and satisfy the estimate

$$0 \leq h(x) \leq Ce^{-\beta|x|}$$

for some $\beta > 0$. Unlike the PFNLS, it is being assumed here that the parametric gain is spatially dependent.

ASSUMPTION 5.8. *The wave perturbs smoothly for $\epsilon > 0$.*

Given that the wave perturbs smoothly, standard perturbation theory reveals that the parameters must satisfy

$$\cos 2\theta = \frac{\gamma}{\mu}\frac{\langle \Phi, \Phi\rangle}{\langle h(x)\Phi, \Phi\rangle} \tag{5.11}$$

and that

$$\partial_\epsilon \mathbf{\Phi} = \begin{bmatrix} \mu \sin 2\theta\, L_r^{-1}(h(x)\Phi) \\ L_i^{-1}([-\gamma + \mu\cos 2\theta\, h(x)]\Phi) \end{bmatrix}.$$

Equation (5.11) and the fact that $h(x)$ is even guarantee that $\partial_\epsilon \mathbf{\Phi}$ is uniformly bounded in x.

Due to the fact that R breaks both the rotation and translation symmetry, it is expected that $E(0) \neq 0$. The calculation of $\partial_\epsilon^2 E(0)$ will be accomplished via an application of Lemma 4.9. Set

$$G(x) = (-\gamma + \mu\cos 2\theta\, h(x))\Phi.$$

A routine calculation shows that

$$\begin{aligned} L_\epsilon &= 8\Phi L_i^{-1}(G)\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} + 8\mu\sin 2\theta\, L_r^{-1}(h(x)\Phi)\begin{bmatrix} 0 & -1 \\ 3 & 0 \end{bmatrix} \\ &\quad -\gamma I_2 + \mu h(x)\begin{bmatrix} \cos 2\theta & -\sin 2\theta \\ -\sin 2\theta & -\cos 2\theta \end{bmatrix}. \end{aligned}$$

Since $h(x)$ being even implies that $\partial_\epsilon \mathbf{\Phi}$ is also even, it is straightforward to see that

$$\langle L_\epsilon \psi_{1,1}, v_2\rangle = \langle L_\epsilon \psi_{1,2}, v_1\rangle = 0.$$

Therefore, upon using the result of Lemma 4.9,

$$\partial_\epsilon^2 E(0) = -2\langle L_\epsilon \psi_{1,1}, v_1\rangle\langle L_\epsilon \psi_{1,2}, v_2\rangle. \tag{5.12}$$

In general, of course, it is impossible to explicitly compute the above quantities without greater knowledge of the function $h(x)$. As such, to facilitate the computation it will be assumed for the rest of this discussion that

$$h(x) = 8\alpha\Phi^2, \tag{5.13}$$

where $\alpha > 0$ is arbitrary. Note that the existence condition then reduces to

$$\cos 2\theta = \frac{\gamma}{\mu}\frac{3}{8\alpha\omega}.$$

For this particular $h(x)$,

$$L_r^{-1}(h(x)\Phi) = \alpha\Phi;$$

therefore,

$$\begin{aligned}\langle L_\epsilon \psi_{1,2}, v_2\rangle &= -\frac{2\mu\sin 2\theta}{\partial_\omega\langle\Phi,\Phi\rangle}(\langle h(x)\Phi,\Phi\rangle + 8\langle \Phi L_r^{-1}(h(x)\Phi),\Phi^2\rangle)\\ &= -\frac{64}{3}\alpha\mu\omega^2\sin 2\theta.\end{aligned}$$

In addition, upon using the fact that

$$(\partial_x\Phi)^2 = \omega\left(\Phi^2 - \frac{2}{\omega}\Phi^4\right),$$

it can be seen that

$$\begin{aligned}\langle L_\epsilon \psi_{1,1}, v_1\rangle &= -\frac{4\mu\sin 2\theta}{\langle\Phi,\Phi\rangle}(\langle h(x),(\partial_x\Phi)^2\rangle - 24\langle \Phi L_r^{-1}(h(x)\Phi),(\partial_x\Phi)^2\rangle)\\ &= \frac{64}{15}\alpha\mu\omega^2\sin 2\theta.\end{aligned}$$

Using (5.12), it is now seen that

$$\partial_\epsilon^2 E(0) = \frac{2^{13}}{3^2\cdot 5}\alpha^2\mu^2\omega^4\sin^2 2\theta. \tag{5.14}$$

LEMMA 5.9. *Consider the NPFNLS. When $h(x)$ is given by* (5.13)*, then the solitary wave is unstable.*

Proof. By (5.14), $E(0) > 0$ for $\epsilon > 0$ sufficiently small. Since $E(\lambda)$ is negative for real λ sufficiently large, there then exists a positive real zero for $E(\lambda)$, and hence a positive real eigenvalue. □

Acknowledgments. I would like to thank Jonathan Rubin for his remarks and suggestions, which helped to clarify the presentation. I would also like to thank Björn Sandstede for several illuminating discussions.

REFERENCES

[1] J. ALEXANDER, R. GARDNER, AND C. JONES, *A topological invariant arising in the stability of travelling waves*, J. Reine Angew. Math., 410 (1990), pp. 167–212.

[2] J. ALEXANDER, M. GRILLAKIS, C. JONES, AND B. SANDSTEDE, *Stability of pulses on optical fibers with phase-sensitive amplifiers*, Z. Angew. Math. Phys., 48 (1997), pp. 175–192.

[3] J. ALEXANDER AND C. JONES, *Existence and stability of asymptotically oscillatory triple pulses*, Z. Angew. Math. Phys., 44 (1993), pp. 189–200.

[4] J. ALEXANDER AND C. JONES, *Existence and stability of asymptotically oscillatory double pulses*, J. Reine Angew. Math., 446 (1994), pp. 49–79.

[5] I. Barashenkov, M. Bogdan, and V. Korobov, *Stability diagram of the phase-locked solitons in the parametrically driven, damped nonlinear Schrödinger equation*, Europhys. Lett., 15 (1991), pp. 113–118.
[6] P. Bates and C. Jones, *Invariant manifolds for semilinear partial differential equations*, Dynam. Report., 2 (1989), pp. 1–38.
[7] A. Bose and C. Jones, *Stability of the in-phase travelling wave solution in a pair of coupled nerve fibres*, Indiana U. Math. J., 44 (1995), pp. 189–220.
[8] R. Gardner and C. Jones, *Travelling waves of a perturbed diffusion equation arising in a phase field model*, Indiana U. Math. J., 38 (1989), pp. 1197–1222.
[9] R. Gardner and C. Jones, *Stability of travelling wave solutions of diffusive predator-prey systems*, Trans. Amer. Math. Soc., 327 (1991), pp. 465–524.
[10] M. Grillakis, *Linearized instability for nonlinear Schrödinger and Klein-Gordon equations*, Comm. Pure Appl. Math., 46 (1988), pp. 747–774.
[11] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer-Verlag, New York, 1981.
[12] C. Jones, *Stability of the travelling wave solutions of the Fitzhugh-Nagumo system*, Trans. Amer. Math. Soc., 286 (1984), pp. 431–469.
[13] C. Jones, *Instability of standing waves for non-linear Schrödinger-type equations*, Ergodic Theory Dynamical Systems, 8 (1988), pp. 119–138.
[14] T. Kapitula, *On the stability of travelling waves in weighted L^∞ spaces*, J. Differential Equations, 112 (1994), pp. 179–215.
[15] T. Kapitula, *Stability criterion for bright solitary waves of the perturbed cubic-quintic Schrödinger equation*, Phys. D, 116 (1998), pp. 95–120.
[16] T. Kapitula and B. Sandstede, *A novel instability mechanism for bright solitary-wave solutions to the cubic-quintic Ginzburg-Landau equation*, J. Opt. Soc. Amer. B, 15 (1998), pp. 2757–2762.
[17] T. Kapitula and B. Sandstede, *Stability of bright solitary wave solutions to perturbed nonlinear Schrödinger equations*, Phys. D, to appear.
[18] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 1980.
[19] T. Kaup, *Closure of the squared Zakharov-Shabat eigenstates*, J. Math. Anal. Appl., 54 (1976), pp. 849–864.
[20] T. Kaup, *A perturbation expansion for the Zakharov-Shabat inverse scattering transform*, SIAM J. Appl. Math., 31 (1976), pp. 121–133.
[21] T. Kaup, *Perturbation theory for solitons in optical fibers*, Phys. Rev. A, 42 (1990), pp. 5689–5694.
[22] D. Kaup and T. Lakoba, *The squared eigenfunctions of the massive Thirring model in laboratory coordinates*, J. Math. Phys., 37 (1976), pp. 308–323.
[23] D. Kaup and T. Lakoba, *Variational method: How it can create false instabilities*, J. Math. Phys., 37 (1996), pp. 3442–3462.
[24] D. Kaup and A. Newell, *Evolution equations, singular dispersion relations, and moving eigenvalues*, Adv. Math., 31 (1979), pp. 67–100.
[25] J. Kutz and W. Kath, *Stability of pulses in nonlinear optical fibers using phase-sensitive amplifiers*, SIAM J. Appl. Math., 56 (1996), pp. 611–626.
[26] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.
[27] R. Pego, P. Smereka, and M. Weinstein, *Oscillatory instability of traveling waves for a KdV-Burgers equation*, Phys. D, 67 (1993), pp. 45–65.
[28] R. Pego, P. Smereka, and M. Weinstein, *Oscillatory instability of solitary waves in a continuum model of lattice vibrations*, Nonlinearity, 6 (1995), pp. 921–941.
[29] R. Pego and M. Weinstein, *Eigenvalues, and instabilities of solitary waves*, Philos. Trans. Roy. Soc. London Ser. A, 340 (1992), pp. 47–94.
[30] R. Pego and M. Weinstein, *Evans' function, and Melnikov's integral, and solitary wave instabilities*, in Differential Equations with Applications to Mathematical Physics, Academic Press, Boston, 1993, pp. 273–286.
[31] R. Pego and M. Weinstein, *Asymptotic stability of solitary waves*, Comm. Math. Phys., 164 (1994), pp. 305–349.
[32] J. Rubin, *Stability and bifurcations of standing pulse solutions to an inhomogeneous reaction-diffusion system*, Proc. Roy. Soc. Edinburgh Sect. A, to appear.
[33] J. Rubin and C. Jones, *Bifurcations and edge oscillations in the semiconductor Fabry-Pérot interferometer*, Opt. Comm., 140 (1997), pp. 93–98.
[34] B. Sandstede, *Stability of multiple-pulse solutions*, Trans. Amer. Math. Soc., 350 (1998), pp. 429–472.

[35] A. Taylor and D. Lay, *Introduction to Functional Analysis*, Wiley, New York, 1980.
[36] M. Weinstein, *Modulational stability of ground states of nonlinear Schrödinger equations*, SIAM J. Math. Anal., 16 (1985), pp. 472–491.
[37] M. Weinstein, *Lyapunov stability of ground states of nonlinear dispersive evolution equations*, Comm. Pure Appl. Math., 39 (1986), pp. 51–68.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 298–325

A SOBOLEV SPACE THEORY OF SPDEs WITH CONSTANT COEFFICIENTS ON A HALF LINE*

N. V. KRYLOV† AND S. V. LOTOTSKY‡

Abstract. Equations of the form $du = (au_{xx} + f_x)\,dt + \sum_k (\sigma^k u_x + g^k)\,dw_t^k$ are considered for $t > 0$ and $x > 0$. The unique solvability of these equations is proved in weighted Sobolev spaces with fractional positive or negative derivatives, summable to the power $p \in [2, \infty)$.

Key words. stochastic partial differential equations, Sobolev spaces with weights

AMS subject classifications. 60H15, 35R60

PII. S0036141097326908

Introduction. We are considering the equation

$$du = (au_{xx} + f_x)\,dt + \sum_{k=1}^{\infty} (\sigma^k u_x + g^k)\,dw_t^k$$

in one space dimension for $x > 0$ and $t > 0$ with some initial condition at $t = 0$ and zero boundary condition at $x = 0$. Here w_t^k are independent one-dimensional Wiener processes and f and g^k are some given functions of (ω, t, x). The functions a and σ^k are assumed to depend only on ω and t. Such equations with a finite number of the processes w_t^k appear, for instance, in nonlinear filtering problems for partially observable diffusions (see [11]). Considering infinitely many w_t^k turns out to be instrumental in treating equations for measure valued processes, for instance, driven by space–time white noise (see [8] or [6]).

Our main goal is to prove solvability of such equations in spaces similar to Sobolev spaces, in which derivatives are understood as generalized functions, the number of derivatives may be fractional or negative, and underlying power of summability is $p \in [2, \infty)$.

The motivation for this goal is explained in detail in [5] or [8], where an L_p-theory is developed for the equations in the whole space. We only mention that if $p = 2$, the theory was developed long ago and an account of it can be found, for instance, in [11]. The case of equations in domains is also treated in [11]. However, the solvability is only proved in spaces W_2^1 of functions having one generalized derivative in x square summable in (ω, t, x). It turns out that going to better smoothness of solutions is not possible in spaces W_2^n and one needs to consider Sobolev spaces with weights, allowing derivatives to blow up near the boundary. The theory of solvability in Hilbert spaces like W_2^n with weights is developed in [1] and [10], where n is an integer. Here we show what happens if one takes a fractional or negative number of derivatives and replaces 2 with any $p \geq 2$. By the way, according to [2], it is not possible to take $p < 2$ when stochastic terms are present in the equation.

*Received by the editors September 8, 1997; accepted for publication (in revised form) March 20, 1998; published electronically November 23, 1998.
http://www.siam.org/journals/sima/30-2/32690.html

†127 Vincent Hall, University of Minnesota, Minneapolis, MN 55455 (krylov@math.umn.edu). The work of this author was partially supported by NSF grant DMS–9625483.

‡Department of Mathematics, M.I.T., Room 2-247, 77 Massachusetts Avenue, Cambridge, MA 02139-4307 (lototsky@math.mit.edu). The work of this author was partially supported by the NSF through the Institute for Mathematics and Its Applications.

Unlike the above mentioned works, we only concentrate on the one-dimensional case. There are several reasons for that, the main being that even in the case of Hilbert spaces in [1] the central estimates are first proved in the one-dimensional case and after this there is still a rather long way to go to get to multidimensional domains. Our treatment of the one-dimensional case is long itself.

One of main difficulties in developing the theory presented below was finding right spaces. The idea was to find a scale of spaces like in [11], [5], or [8] generated by fractional powers of a certain operator, which is $1-\Delta$ in [11], [5], and [8]. From the results of [1] and [10] one can guess that $xD = x\partial/\partial x$ should be such an operator in our case. Elliptic second-order operators are more appropriate if one wants to define fractional powers and expects them to have nice properties. Therefore, our first attempt was to try the operator $L = xD(xD) + xD - c$, which is formally self-adjoint for any constant c. However, after having constructed the theory we noticed that the same spaces can be defined as images of spaces from [5] or [8] under certain linear mapping. This made using the results from [5] and [8] easier and allowed us to avoid developing solvability theory for L and investigating the semigroup and the resolvent associated with this operator.

In [11], [5], and [8] the solution is sought for in the same scale of spaces (at least as far as the space variables are concerned) as the one to which the free terms f and g belong. Surprisingly enough this is not the case in our situation, and this causes many difficulties practically at each step. The origin of all unusual features of our theory lies in the fact that there are no operators commuting with $\partial/\partial x$ and generating our scale of spaces. To give one more example of what is unusual we state the following theorem, which can be obtained from Theorem 3.2 after changing variables $v(t,x) = e^{x(\alpha-1)}u(t,e^x)$, where $\alpha = \theta/p$.

THEOREM 0.1. *Let $\alpha \in (0,1)$, $p \in (1,\infty)$, $T \in (0,\infty]$, and $f \in L_p([0,\infty)\times\mathbb{R})$. Then in the class of functions $v(t,x)$, $t \in [0,T]$, $x \in \mathbb{R}$ such that*

$$\int_0^T \int_{\mathbb{R}} [|v_x|^p + |v|^p]\,dxdt < \infty,$$

the equation

$$e^{2x}v_t = v_{xx} + (1-2\alpha)v_x - (1-\alpha)\alpha v + f_x \tag{0.1}$$

on $(0,T)\times\mathbb{R}$ with zero initial condition has a unique solution. In addition, this solution satisfies

$$\int_0^T \int_{\mathbb{R}} [|v_x|^p + |v|^p]\,dxdt \le N(\alpha,p)\int_0^T \int_{\mathbb{R}} |f|^p\,dxdt.$$

Surprising in this theorem is that if we replace e^{2x} with 1 in (0.1), then the result becomes well known and is true for any finite T (now with N depending on T too). The presence of e^{2x} makes (0.1) degenerate, and usually results for degenerate equations differ very much from those for nondegenerate cases. Actually, we do not know much about (0.1). In particular, it would be interesting to know whether Theorem 0.1 remains true if we replace the term $(1-2\alpha)v_x$ in (0.1) with bv_x where b is an arbitrary constant.

The article is organized as follows. In section 1 we introduce and investigate basic spaces with weights of functions of $x \in (0,\infty)$. Section 2 is devoted to stochastic Banach spaces of functions of (ω,t,x) satisfying zero boundary condition at $x = 0$.

This condition is expressed by means of requirement (2.1). In section 3 we prove our main Theorem 3.2 about unique solvability of our equations. The reader will see the very core of our technique in the proof of Lemma 3.6. Rather long section 4 contains the proof of the main particular case of Theorem 3.2, which is stated as Lemma 3.5.

1. Sobolev spaces with weights. For $\gamma \in \mathbb{R}$ and $p \in (1, \infty)$ let $H_p^\gamma = H_p^\gamma(\mathbb{R})$ be the spaces of Bessel potentials (see, for instance, [13]) which are formally given by $H_p^\gamma = \Lambda^{-\gamma} L_p(\mathbb{R})$, where $\Lambda := (1 - D^2)^{1/2}$ and $D = d/dx$. One knows that the elements of H_p^γ are distributions and $C_0^\infty = C_0^\infty(\mathbb{R})$ is dense in H_p^γ. Let $\mathcal{D}(\mathbb{R})$ and $\mathcal{D}(\mathbb{R}_+)$ be the sets of all distributions on $C_0^\infty(\mathbb{R})$ and $C_0^\infty(\mathbb{R}_+)$, respectively, where $\mathbb{R}_+ = (0, \infty)$. If $f \in \mathcal{D}(\mathbb{R}_+)$ and $\theta \in \mathbb{R}$, then the expression $h(x) := f(e^x)e^{x\theta/p}$ is well defined and is a distribution on $\mathbb{R}$. Indeed, the action of h on a test function $\phi \in C_0^\infty(\mathbb{R})$ is defined as $(h, \phi) = (f, \psi)$, where $\psi(x) := \phi(\log x)x^{\theta/p-1}$. We denote $h = Q_{p,\theta} f$ in this way defining a one-to-one operator

$$Q_{p,\theta} : f(x) \to f(e^x)e^{x\theta/p}.$$

Definition 1.1. *We write $f \in H_{p,\theta}^\gamma$ $(= H_{p,\theta}^\gamma(\mathbb{R}_+))$ if and only if $Q_{p,\theta} f = h \in H_p^\gamma$. We write $L_{p,\theta} = H_{p,\theta}^0$. For $f \in H_{p,\theta}^\gamma$ we define*

$$||f||_{H_{p,\theta}^\gamma} = ||Q_{p,\theta} f||_{H_p^\gamma}.$$

Remark 1.2. Since H_p^γ is a Banach space, so is $H_{p,\theta}^\gamma$ with the norm introduced above. Also since $C_0^\infty(\mathbb{R})$ is dense in H_p^γ, the set $C_0^\infty(\mathbb{R}_+)$ is dense in $H_{p,\theta}^\gamma$.

Remark 1.3. Define $\Lambda_{p,\theta}^\gamma = Q_{p,\theta}^{-1} \Lambda^\gamma Q_{p,\theta}$. Then for any $\gamma, \mu, \theta \in \mathbb{R}$ the operator $\Lambda_{p,\theta}^\gamma$ is an isometric operator from $H_{p,\theta}^\mu$ onto $H_{p,\theta}^{\mu-\gamma}$.

Indeed, by definition,

$$||\Lambda_{p,\theta}^\gamma u||_{H_{p,\theta}^{\mu-\gamma}} = ||Q_{p,\theta}\Lambda_{p,\theta}^\gamma u||_{H_p^{\mu-\gamma}} = ||\Lambda^\gamma Q_{p,\theta} u||_{H_p^{\mu-\gamma}}$$

$$= ||Q_{p,\theta} u||_{H_p^\mu} = ||u||_{H_{p,\theta}^\mu}.$$

Remark 1.4. The norm in $H_{p,\theta}^\gamma$ contains norms of, so to speak, γ derivatives of u. However, it scales in the same way for any γ. We mean that, due to translation invariance of norms in H_p^γ, for any constant $a > 0$ and $u \in H_{p,\theta}^\gamma$,

$$||u(a\,\cdot)||_{H_{p,\theta}^\gamma}^p = a^{-\theta}||u||_{H_{p,\theta}^\gamma}^p.$$

Remark 1.5. Define M as the operator of multiplying by x, $M : u(x) \to xu(x)$. It turns out that for any $\gamma \in \mathbb{R}$ the operator MD is a bounded operator from $H_{p,\theta}^\gamma$ into $H_{p,\theta}^{\gamma-1}$ and if, in addition, $\theta \neq 0$, then MD maps $H_{p,\theta}^\gamma$ onto $H_{p,\theta}^{\gamma-1}$ and its inverse is also bounded.

Indeed, an easy computation shows that

$$Q_{p,\theta} MDu = LQ_{p,\theta} u, \quad MDu = Q_{p,\theta}^{-1} LQ_{p,\theta} u,$$

where $Lv = Dv - v\theta/p$. One knows (see, for instance, p. 263 in [12]) that for any constant ν the operator $v \to Dv + \nu v$ is a bounded operator from H_p^γ into $H_p^{\gamma-1}$ and if ν is real and $\nu \neq 0$, then it maps H_p^γ onto $H_p^{\gamma-1}$ and its inverse is bounded. This and the definition of $H_{p,\theta}^\gamma$ obviously imply our assertion.

Remark 1.6. Functions in $H^\gamma_{p,\theta}$ are different from those in H^γ_p only in what concerns their behavior near zero and infinity. More precisely, if $[a,b] \subset \mathbb{R}_+$ and $f = 0$ outside $[a,b]$, then by the results on changing variables and pointwise multipliers (see Theorem 4.3.2 and Corollary 4.2.2 of [13]) $||f||_{H^\gamma_{p,\theta}} \le N||f||_{H^\gamma_p} \le N||f||_{H^\gamma_{p,\theta}}$, where N is independent of f.

It is convenient here also to notice that for the same f we have

$$||f||_{H^\gamma_p} \le N||Df||_{H^{\gamma-1}_p} \le N||f||_{H^\gamma_p},$$

with N independent of f.

Indeed, the inequality on the right is known to be true even for any $f \in H^\gamma_p$. As far as the left inequality is concerned, by Remark 1.5 we have

$$||f||_{H^\gamma_p} \le N||f||_{H^\gamma_{p,1}} \le N||MDf||_{H^{\gamma-1}_{p,1}} \le N||\eta Df||_{H^{\gamma-1}_p},$$

where $\eta \in C_0^\infty(\mathbb{R})$ and $\eta(x) = x$ on $[a,b]$. It only remains to remember (see [13]) that such η is a pointwise multiplier in any space $H^{\gamma-1}_p$.

Remark 1.7. Upon noticing that $DMu = MDu+u$, as in Remark 1.5 we conclude that for any $\gamma \in \mathbb{R}$ the operator DM is a bounded operator from $H^\gamma_{p,\theta}$ into $H^{\gamma-1}_{p,\theta}$ and if, in addition, $\theta \ne p$, then DM maps $H^\gamma_{p,\theta}$ onto $H^{\gamma-1}_{p,\theta}$ and its inverse is also bounded.

Remark 1.8. Let $\theta \ne 0$, $u \in \bigcup_\mu H^\mu_{p,\theta}$, and $MDu \in H^\gamma_{p,\theta}$. Then $u \in H^{\gamma+1}_{p,\theta}$ and $||u||_{H^{\gamma+1}_{p,\theta}} \le N||MDu||_{H^\gamma_{p,\theta}}$.

Indeed, by Remark 1.5 there is $v \in H^{\gamma+1}_{p,\theta}$ such that $MDv = MDu$ and $||v||_{H^{\gamma+1}_{p,\theta}} \le N||MDu||_{H^\gamma_{p,\theta}}$. Then $v' = u'$ and $v-u = c$, where c is a constant. Since $v, u \in H^\mu_{p,\theta}$ for some μ, we have $c \in H^\mu_{p,\theta}$, which is only possible if $c = 0$. Therefore, $u = v \in H^{\gamma+1}_{p,\theta}$.

Remark 1.9. Let $\theta \ne p$, $u \in \bigcup_\mu H^\mu_{p,\theta}$, and $DMu \in H^\gamma_{p,\theta}$. Then $u \in H^{\gamma+1}_{p,\theta}$ and $||u||_{H^{\gamma+1}_{p,\theta}} \le N||DMu||_{H^\gamma_{p,\theta}}$.

Indeed, one can repeat the argument in Remark 1.8 relying on Remark 1.7 instead of Remark 1.5 and noticing that from the equality $DMv = DMu$ it follows that $v - u = c/x$, where c is a constant.

Remark 1.5 and the observation that $H^0_{p,\theta} = L_{p,\theta}$ is just an L_p-space of functions on $\mathbb{R}_+$ with measure $m_\theta(dx) = x^{\theta-1}\,dx$ yield inequalities (1.1) in the following useful result, which can also be restated in a natural way on the basis of Remark 1.7.

THEOREM 1.10. *If γ is an integer satisfying $\gamma \ge 1$ and $\theta \ne 0$, then for any $u \in H^\gamma_{p,\theta}$ we have*

$$||(MD)^\gamma u||_{L_p(\mathbb{R}_+,m_\theta)} \le N||u||_{H^\gamma_{p,\theta}} \le N||(MD)^\gamma u||_{L_p(\mathbb{R}_+,m_\theta)}, \tag{1.1}$$

$$\sum_{n=1}^\gamma ||M^n D^n u||_{L_p(\mathbb{R}_+,m_\theta)} \le N||u||_{H^\gamma_{p,\theta}} \le N\sum_{n=1}^\gamma ||M^n D^n u||_{L_p(\mathbb{R}_+,m_\theta)}, \tag{1.2}$$

where N is independent of u. Thus, the space $H^\gamma_{p,\theta}$ can also be defined as a closure of the set $C_0^\infty(\mathbb{R}_+)$ with respect to either of the norms

$$||(MD)^\gamma \cdot||_{L_p(\mathbb{R}_+,m_\theta)}, \qquad \sum_{n=1}^\gamma ||M^n D^n \cdot||_{L_p(\mathbb{R}_+,m_\theta)}.$$

To prove (1.2) observe that for any integer $k \geq 1$,

$$(MD)^k = \sum_{n=1}^{k} c^{k,n} M^n D^n, \tag{1.3}$$

where $c^{k,n}$ are some constants and $c^{k,k} = 1$. This and the inequality on the right in (1.1) give us the inequality on the right in (1.2). On the other hand, one can solve the triangular system (1.3) with respect to $M^n D^n$. Then from the inequality on the left in (1.1) we get

$$\sum_{n=1}^{\gamma} ||M^n D^n u||_{L_p(\mathbb{R}_+, m_\theta)} \leq N \sum_{n=1}^{\gamma} ||(MD)^n u||_{L_p(\mathbb{R}_+, m_\theta)}$$

$$\leq N \sum_{n=1}^{\gamma} ||u||_{H^n_{p,\theta}} \leq N ||u||_{H^\gamma_{p,\theta}},$$

which proves the inequality on the left in (1.2).

The following theorem will play the most important role in obtaining results for equations on $\mathbb{R}_+$ from those on $\mathbb{R}$.

THEOREM 1.11. *Let $\zeta \in C_0^\infty(\mathbb{R}_+)$, $\gamma, \theta \in \mathbb{R}$, and $p \in (1, \infty)$. Then there exists a constant N depending only on ζ, γ, p, and θ such that, for any $u \in H^\gamma_{p,\theta}$,*

$$\sum_{n=-\infty}^{\infty} e^{n\theta} ||\zeta u(e^n \cdot)||^p_{H^\gamma_p} \leq N ||u||^p_{H^\gamma_{p,\theta}}.$$

In addition, if there is a $\delta > 0$ such that

$$\sum_{n=-\infty}^{\infty} e^{(n-x)\theta} |\zeta(e^{x-n})|^p \geq \delta \tag{1.4}$$

for all $x \in [0,1]$, then

$$||u||^p_{H^\gamma_{p,\theta}} \leq N \sum_{n=-\infty}^{\infty} e^{n\theta} ||\zeta u(e^n \cdot)||^p_{H^\gamma_p},$$

where N depends on δ as well.

Proof. Since the functions $\zeta(x) u(e^n x)$ vanish outside the support of ζ, by the change of variables (see Theorem 4.3.2 in [13])

$$e^{n\theta} ||\zeta u(e^n \cdot)||^p_{H^\gamma_p} \leq N e^{n\theta} ||\zeta(e^\cdot) u(e^{\cdot + n})||^p_{H^\gamma_p}$$

with N independent of n, u. By translation invariance of the norm in H^γ_p the last expression equals

$$e^{n\theta} ||\zeta(e^{\cdot - n}) u(e^\cdot)||^p_{H^\gamma_p} = ||\eta(e^{\cdot - n}) Q_{p,\theta} u||^p_{H^\gamma_p},$$

where $\eta(e^{x-n}) = \zeta(e^{x-n}) e^{(n-x)\theta/p}$. Next it is easy to find a finite m such that for $x \in [0,1]$,

$$I(x) := \sum_{n=-\infty}^{\infty} |\eta(e^{x-n})|^p = \sum_{n=-\infty}^{\infty} |\zeta(e^{x-n})|^p e^{(x-n)\theta}$$

$$= \sum_{|n| \leq m} |\zeta(e^{x-n})|^p e^{(x-n)\theta}.$$

It follows that $I(x)$ is bounded on $[0,1]$. On the other hand $I(x)$ is obviously periodic with period 1. Thus $I(x)$ is bounded on $\mathbb{R}$. The same is true for

$$\sum_{n=-\infty}^{\infty} |(\eta(e^{x-n}))'|^p, \quad \sum_{n=-\infty}^{\infty} |(\eta(e^{x-n}))''|^p,$$

and so on. By Theorem 2.2 and Remark 2.1 of [2]

$$\sum_{n=-\infty}^{\infty} ||\eta(e^{\cdot-n})Q_{p,\theta}u||^p_{H^\gamma_p} \leq N||Q_{p,\theta}u||^p_{H^\gamma_p},$$

which yields our first assertion.

To prove the second one we use the same resources as above and get

$$||Q_{p,\theta}u||^p_{H^\gamma_p} \leq N\sum_{n=-\infty}^{\infty} ||\eta(e^{\cdot-n})Q_{p,\theta}u||^p_{H^\gamma_p} = N\sum_{n=-\infty}^{\infty} e^{n\theta}||\zeta(e^{\cdot-n})u(e^{\cdot})||^p_{H^\gamma_p}$$

$$= N\sum_{n=-\infty}^{\infty} e^{n\theta}||\zeta(e^{\cdot})u(e^{\cdot+n})||^p_{H^\gamma_p} \leq N\sum_{n=-\infty}^{\infty} e^{n\theta}||\zeta u(e^n\,\cdot)||^p_{H^\gamma_p}.$$

The theorem is proved.

Remark 1.12. Similar to properties of $I(x)$ in the above proof, we find that if $\zeta \in C_0^\infty(\mathbb{R}_+)$ and $\beta \in \mathbb{R}$, then $\sum_n e^{(n+x)\beta}\zeta(e^{n+x})$ is bounded on $\mathbb{R}$, which after substituting $\log x$ in place of x implies that $\sum_n e^{n\beta}\zeta(e^n x) \leq Nx^{-\beta}$ on $\mathbb{R}_+$.

The following theorem is used in establishing some properties of our stochastic Banach spaces.

THEOREM 1.13. *Recall that the operator M is defined by $Mu(x) = xu(x)$ and let $\theta, \gamma \in \mathbb{R}$, $\theta \neq p$. Then*

$$M^{-1}u \in H^{\gamma+1}_{p,\theta} \iff Du \in H^\gamma_{p,\theta} \text{ and } M^{-1}u \in \bigcup_\mu H^\mu_{p,\theta}. \tag{1.5}$$

In addition, under either one of the above conditions

$$||M^{-1}u||_{H^{\gamma+1}_{p,\theta}} \leq N||Du||_{H^\gamma_{p,\theta}} \leq N||M^{-1}u||_{H^{\gamma+1}_{p,\theta}}. \tag{1.6}$$

Proof. If $M^{-1}u \in H^{\gamma+1}_{p,\theta}$, then by Remark 1.7 we have

$$Du = DM(M^{-1}u) \in H^\gamma_{p,\theta}$$

and the right inequality in (1.6) holds. On the other hand, under the condition on the right in (1.5) we have

$$DM(M^{-1}u) \in H^\gamma_{p,\theta} \quad \text{and} \quad M^{-1}u \in \bigcup_\mu H^\mu_{p,\theta},$$

which by Remark 1.9 yields $M^{-1}u \in H^{\gamma+1}_{p,\theta}$ and the inequality on the left in (1.6). The theorem is proved.

The following result will also be used in the future.

LEMMA 1.14. *For any constants p, θ, α we have*

$$
\begin{gathered}
Q_{p,\theta}^{-1} D Q_{p,\theta} = bI + MD, \quad Q_{p,\theta}^{-1} D^2 Q_{p,\theta} = (bI + MD)^2, \quad DM = MD + I, \\
M^\alpha \Lambda_{p,\theta}^2 M^{-\alpha} = \Lambda_{p,\theta}^2 + c_1 I + c_2 MD, \quad M\Lambda_{p,\theta}^2 - \Lambda_{p,\theta}^2 M = MP_1, \\
\Lambda_{p,\theta}^2 D - D\Lambda_{p,\theta}^2 = P_1 D, \\
\Lambda_{p,\theta}^2 DM - D\Lambda_{p,\theta}^2 M = P_1 DM, \quad \Lambda_{p,\theta}^2 D^2 M - D^2 \Lambda_{p,\theta}^2 M = 4DP_2,
\end{gathered} \tag{1.7}
$$

where $b = \theta/p$, I is the identity operator, c_i are certain constants, and

$$
P_1 := (2b+1)I + 2MD, \quad P_2 := bDM + (MD)(DM).
$$

Furthermore, for any $\theta, \gamma \in \mathbb{R}$ there exists a constant $N = N(\gamma, \theta, p)$ such that for any $u \in H_{p,\theta}^{\gamma+2}$,

$$
||P_1 u||_{H_{p,\theta}^{\gamma+1}} + ||P_2 u||_{H_{p,\theta}^{\gamma}} \le N ||u||_{H_{p,\theta}^{\gamma+2}}. \tag{1.8}
$$

Indeed, equalities (1.7) are checked out by straightforward computations and (1.8) follows immediately from Remarks 1.5 and 1.7.

2. Stochastic Banach spaces on $\mathbb{R}_+$. Let $(\Omega, \mathcal{F}, P)$ be a complete probability space, $(\mathcal{F}_t, t \ge 0)$ be an increasing filtration of σ-fields $\mathcal{F}_t \subset \mathcal{F}$ containing all P-null subsets of Ω, and $\mathcal{P}$ be the predictable σ-field generated by $(\mathcal{F}_t, t \ge 0)$. Let $\{w_t^k; k = 1, 2, ...\}$ be a family of independent one-dimensional $\mathcal{F}_t$-adapted Wiener processes defined on $(\Omega, \mathcal{F}, P)$. We are going to use the Banach spaces $\mathbb{H}_p^\gamma(\tau)$, $\mathbb{H}_p^\gamma(\tau, l_2)$, and $\mathcal{H}_p^\gamma(\tau)$ introduced in [5] or [8], where we take $d = 1$. Also throughout the remaining part of the paper $\theta \ne 0$, $\theta \ne p$, and $p \ge 2$ unless another range of p is specified explicitly.

DEFINITION 2.1. *Let τ be a stopping time, f and g^k, $k = 1, 2...$, be $\mathcal{D}(\mathbb{R}_+)$-valued $\mathcal{P}$-measurable functions defined on $(\!(0, \tau]\!]$. We write $f \in \mathbb{H}_{p,\theta}^\gamma(\tau)$ and $g \in \mathbb{H}_{p,\theta}^\gamma(\tau, l_2)$ if and only if $Q_{p,\theta} f \in \mathbb{H}_p^\gamma(\tau)$ and $Q_{p,\theta} g \in \mathbb{H}_p^\gamma(\tau, l_2)$, respectively. We also denote*

$$
||f||_{\mathbb{H}_{p,\theta}^\gamma(\tau)} = ||Q_{p,\theta} f||_{\mathbb{H}_p^\gamma(\tau)}, \quad ||g||_{\mathbb{H}_{p,\theta}^\gamma(\tau, l_2)} = ||Q_{p,\theta} g||_{\mathbb{H}_p^\gamma(\tau, l_2)},
$$

$$
\mathbb{H}_{p,\theta}^\gamma = \mathbb{H}_{p,\theta}^\gamma(\infty), \quad \mathbb{H}_{p,\theta}^\gamma(l_2) = \mathbb{H}_{p,\theta}^\gamma(\infty, l_2), \quad \mathbb{L}_{...,...} = \mathbb{H}_{...,...}^0.
$$

In the case $f \in \mathbb{H}_{p,\theta}^\gamma(\tau)$, $g \in \mathbb{H}_{p,\theta}^{\gamma+1}(\tau, l_2)$ we write $(f, g) \in \mathcal{F}_{p,\theta}^\gamma(\tau)$ and

$$
||(f, g)||_{\mathcal{F}_{p,\theta}^\gamma(\tau)} = ||f||_{\mathbb{H}_{p,\theta}^\gamma(\tau)} + ||g||_{\mathbb{H}_{p,\theta}^{\gamma+1}(\tau, l_2)}.
$$

Finally, we introduce spaces of initial data. We write $u_0 \in U_{p,\theta}^\gamma$ if and only if $M^{2/p-1} u_0 \in L_p(\Omega, \mathcal{F}_0, H_{p,\theta}^{\gamma-2/p})$ and denote

$$
||u_0||_{U_{p,\theta}^\gamma}^p = E||M^{2/p-1} u_0||_{H_{p,\theta}^{\gamma-2/p}}^p.
$$

DEFINITION 2.2. *For a $\mathcal{D}(\mathbb{R}_+)$-valued function u defined on $\Omega \times [0, \infty)$ with $u(0, \cdot) \in U_{p,\theta}^{\gamma+1}$ and*

$$
M^{-1} u \in \bigcup_\mu \bigcap_{T>0} \mathbb{H}_{p,\theta}^\mu(\tau \wedge T), \tag{2.1}
$$

we write $u \in \mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)$ *if and only if* $u_x \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ *and there exists* $(f,g) \in \mathcal{F}^{\gamma-1}_{p,\theta}(\tau)$ *such that for any* $\phi \in C_0^\infty(\mathbb{R}_+)$ *we have*

$$(u(t,\cdot),\phi) = (u(0,\cdot),\phi) + \int_0^t (M^{-1}f(s,\cdot),\phi)\,ds + \sum_{k=1}^{\infty}\int_0^t (g^k(s,\cdot),\phi)\,dw^k_s \tag{2.2}$$

for all $t \le \tau$ *at once with probability one. In this situation we also write* $M^{-1}f = \tilde{\mathbb{D}}u$, $g = \tilde{\mathbb{S}}u$,

$$du = M^{-1}f\,dt + g^k\,dw^k_t$$

and define $\mathfrak{H}^{\gamma+1}_{p,\theta,0}(\tau) = \mathfrak{H}^{\gamma+1}_{p,\theta}(\tau) \cap \{u : u(0,\cdot) = 0\}$,

$$||u||^p_{\mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)} = ||u_x||^p_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} + ||(f,g)||^p_{\mathcal{F}^{\gamma-1}_{p,\theta}(\tau)} + ||u(0,\cdot)||^p_{U^{\gamma+1}_{p,\theta}}. \tag{2.3}$$

As always, we drop τ *in* $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ *and* $\mathcal{F}^{\gamma}_{p,\theta}(\tau)$ *if* $\tau = \infty$.

Remark 2.3 (cf. Remark 3.3 in [8]). Given $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, there exists only one pair of functions f and g in Definition 2.2. Therefore, the notation $M^{-1}f = \tilde{\mathbb{D}}u$, $g = \tilde{\mathbb{S}}u$, and (2.3) make sense.

It is also worth noting that the last series in (2.2) converges uniformly in t on each interval $[0, \tau \wedge T]$, $T \in (0,\infty)$, in probability.

Remark 2.4. It follows from Theorem 1.13 that, in Definition 2.2, the two requirements (2.1) and $u_x \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ can be replaced with only one: $M^{-1}u \in \mathbb{H}^{\gamma+1}_{p,\theta}(\tau)$. In addition,

$$||M^{-1}u||_{\mathbb{H}^{\gamma+1}_{p,\theta}(\tau)} \le N||u_x||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} \le N||M^{-1}u||_{\mathbb{H}^{\gamma+1}_{p,\theta}(\tau)},$$

where $N = N(\gamma,\theta,p)$.

Remark 2.5. The space $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ is *not* $Q^{-1}_{p,\theta}\mathcal{H}^{\gamma}_p(\tau)$. However, obviously ϕu lies in $Q^{-1}_{p,\theta}\mathcal{H}^{\gamma}_p(\tau)$ for any $\phi \in C_0^\infty(\mathbb{R}_+)$ if $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$. By Theorem 3.7 of [8] this easily implies that if $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ and $||u||_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)} = 0$, then u is indistinguishable from zero.

Of course, we identify elements of $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ which are indistinguishable.

Remark 2.6. The spaces $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ and $\mathfrak{H}^{\gamma}_{p,\theta,0}(\tau)$ are Banach spaces.

Indeed, their completeness is obtained as follows. If u_n is a Cauchy sequence in $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, then $M^{-1}u_n$ is a Cauchy sequence in $\mathbb{H}^{\gamma}_{p,\theta}(\tau)$ by Remark 2.4 and hence it converges to some $M^{-1}u \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$. Also, $M\tilde{\mathbb{D}}u_n \to f$ and $\tilde{\mathbb{S}}u_n \to g$ for some $(f,g) \in \mathcal{F}^{\gamma-2}_{p,\theta}(\tau)$.

Next, for any $\phi \in C_0^\infty(\mathbb{R}_+)$ the sequence ϕu_n is a Cauchy sequence in $\mathcal{H}^{\gamma}_p(\tau)$, which is a Banach space by Theorem 3.7 of [8]. This easily implies that u has a modification $\bar{u}$ such that $\phi\bar{u}$ belongs to $\mathcal{H}^{\gamma}_p(\tau)$ for any $\phi \in C_0^\infty(\mathbb{R}_+)$, and $\bar{u}$ satisfies (2.2), so that $\bar{u} \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$. One treats $\mathfrak{H}^{\gamma}_{p,\theta,0}(\tau)$ similarly.

Remark 2.7. By Remark 1.5 it follows that $f \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$ if and only if there exists a unique $h \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ such that $M^{-1}f = Dh$. In addition, the norms of f and h are equivalent. Hence, one obtains the same space $\mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)$ if in Definition 2.2 one replaces $M^{-1}f$ with f_x and instead of the condition $(f,g) \in \mathcal{F}^{\gamma-1}(\tau)$ requires $f \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$, $g \in \mathbb{H}^{\gamma}_{p,\theta}(\tau, l_2)$. In this case one obtains an equivalent norm by replacing $||(f,g)||^p_{\mathcal{F}^{\gamma-1}_{p,\theta}(\tau)}$ in (2.3) with

$$||f||^p_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} + ||g||^p_{\mathbb{H}^{\gamma}_{p,\theta}(\tau,l_2)}.$$

Remark 2.8. If $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, then $v := MDu \in \mathfrak{H}^{\gamma-1}_{p,\theta}(\tau)$ and

$$||MDu||_{\mathfrak{H}^{\gamma-1}_{p,\theta}(\tau)} \le N(\gamma,\theta,p)||u||_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}.$$

Indeed, we have $M^{-1}v = Du \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$, which by Remark 2.4 gives us a part of the needed properties of v. Also by Remark 2.7, $du = f_x\,dt + g^k\,dw^k_t$ with $f \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau)$ and $g \in \mathbb{H}^{\gamma-1}_{p,\theta}(\tau, l_2)$, so that $dv = (MDf - f)_x\,dt + MDg^k\,dw^k_t$, where by Remark 1.5

$$||MDf - f||_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)} \le N||f||_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)}, \quad ||MDg||_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau,l_2)} \le N||g||_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau,l_2)},$$

$$||M^{2/p-1}v(0,\cdot)||_{H^{\gamma-1-2/p}_{p,\theta}} = ||MD(M^{2/p-1}u(0,\cdot)) - (2/p-1)M^{2/p-1}u(0,\cdot)||_{H^{\gamma-1-2/p}_{p,\theta}}$$
$$\le N||M^{2/p-1}u(0,\cdot)||_{H^{\gamma-2/p}_{p,\theta}}.$$

Remark 2.9. From Remark 1.3 we have

$$||\Lambda^{\gamma}_{p,\theta}u||_{\mathbb{H}^{\mu}_{p,\theta}(\tau)} = ||u||_{\mathbb{H}^{\mu+\gamma}_{p,\theta}(\tau)}.$$

The assertions of the following theorem are straightforward corollaries of Remark 2.4 and of two Sobolev theorems. One says that $H^{\gamma}_p \subset \mathcal{C}^{\delta}$ if $\delta := \gamma - 1/p > 0$, where $\mathcal{C}^{\delta} = \mathcal{C}^{\delta}(\mathbb{R})$ is the Zygmund space (which differs from the usual Hölder space $C^{\delta} = C^{\delta}(\mathbb{R})$ only if δ is an integer; see [13]). The second one says that $H^{\gamma}_p \subset H^{\mu}_q$ if $\mu < \gamma$ and $\gamma - 1/p = \mu - 1/q$. These theorems are easily rewritten in terms of our spaces $H^{\gamma}_{p,\theta} = Q^{-1}_{p,\theta}H^{\gamma}_p$.

THEOREM 2.10. (i) *If $\alpha := \gamma - 1/p > 0$ and $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, then $Q_{p,\theta}M^{-1}u \in L_p(\langle\!\langle 0,\tau]\!],\mathcal{C}^{\alpha})$, where $\mathcal{C}^{\alpha}$ is the Zygmund space. In addition,*

$$E\int_0^{\tau} ||Q_{p,\theta}M^{-1}u(t,\cdot)||^p_{\mathcal{C}^{\alpha}}\,dt \le N(d,\gamma,p)||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}.$$

(ii) *If $\mu < \gamma$, $\gamma - 1/p = \mu - 1/q$, and $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$, then*

$$E\int_0^{\tau} ||M^{-1}u(t,\cdot)||^p_{H^{\mu}_{q,\theta q/p}}\,dt \le N(d,\gamma,\mu,p)||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}.$$

In order to prove the solvability even of the simplest equations we need the following embedding theorem. However, the way in which the right-hand side of (2.4) depends on T will not be used.

THEOREM 2.11. *Let $T \in (0,\infty)$ be a constant and let $\tau \le T$. Then for any function $u \in \mathfrak{H}^{\gamma}_{p,\theta,0}(\tau)$, we have*

$$E\sup_{t\le\tau} ||u(t,\cdot)||^p_{H^{\gamma-1}_{p,\theta}} \le N(p,\theta,\gamma)T^{(p-2)/p}||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}. \tag{2.4}$$

To prove this theorem we use the following fact, which is similar to Remark 2.2 of [5] or Remark 4.11 of [8].

LEMMA 2.12. *Let $T \in (0,\infty)$ be a constant and let $\tau \le T$. Let $u \in \mathcal{H}^{\gamma}_{p,0}(\tau)$ and $du = f\,dt + g^k\,dw^k_t$. Then for any constant $c > 0$,*

$$E\sup_{t\le\tau} ||u_x(t,\cdot)||^p_{H^{\gamma-2}_p} \le N(p)T^{(p-2)/2}(c||u_{xx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau)}$$
$$+c^{-1}||f||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} + ||g_x||^p_{\mathbb{H}^{\gamma-2}_p(\tau,l_2)}). \tag{2.5}$$

Proof. As always, it suffices to prove (2.5) for any particular γ and $\tau = T$ (regarding τ see, for instance, the proof of Theorem 7.1 in [8]). We take $\gamma = 2$. Then (2.5) becomes

$$E\sup_{t\le T}||u_x(t,\cdot)||^p_{L_p} \le N(p)T^{(p-2)/2}(c||u_{xx}||^p_{\mathbb{L}_p(T)}$$

$$+c^{-1}||f||^p_{\mathbb{L}_p(T)} + ||g_x||^p_{\mathbb{L}_p(T,l_2)}). \tag{2.6}$$

It suffices to prove this inequality for $c = 1$. Indeed, for any constant $a > 0$ we have $du(t,ax) = f(t,ax)\,dt + g^k(t,ax)\,dw^k_t$ and if (2.6) holds with $c = 1$, then

$$a^{p-1}E\sup_{t\le T}||u_x(t,\cdot)||^p_{L_p} = E\sup_{t\le T}||(u(t,a\,\cdot))_x||^p_{L_p}$$

$$\le NT^{(p-2)/2}(||(u(\cdot,a\,\cdot))_{xx}||^p_{\mathbb{L}_p(T)} + ||f(\cdot,a\,\cdot)||^p_{\mathbb{L}_p(T)} + ||(g(\cdot,a\,\cdot))_x||^p_{\mathbb{L}_p(T,l_2)})$$

$$= NT^{(p-2)/2}(a^{2p-1}||u_{xx}||^p_{\mathbb{L}_p(T)} + a^{-1}||f||^p_{\mathbb{L}_p(T)} + a^{p-1}||g_x||^p_{\mathbb{L}_p(T,l_2)}).$$

This proves (2.6) with a^p in place of c.

We further transform (2.6) with $c = 1$ by denoting $v = u_x$ and $h^k = g^k_x$, so that $dv = f_x\,dt + h^k\,dw^k_t$ and $v \in \mathcal{H}^1_{p,0}(T)$. We see that we only need to prove that

$$E\sup_{t\le T}||v(t,\cdot)||^p_{L_p} \le N(p)T^{(p-2)/2}(||v_x||^p_{\mathbb{L}_p(T)}$$

$$+||f||^p_{\mathbb{L}_p(T)} + ||h||^p_{\mathbb{L}_p(T,l_2)}). \tag{2.7}$$

By Theorem 2.1 of [5] or Theorem 4.10 of [8] and by the observation that $dv = (v_{xx} + (f - v_x)_x)\,dt + h^k\,dw^k_t$, for any $\lambda, T > 0$ we have

$$E\sup_{t\le T}(e^{-p\lambda t}||v(t,\cdot)||^p_{L_p}) \le N(||e^{-\lambda t}\bar f||^p_{\mathbb{L}_p(T)} + ||e^{-\lambda t}h||^p_{\mathbb{L}_p(T,l_2)}),$$

where $N = N(p,\lambda)$ and $\bar f = f - v_x$. For $\lambda = 1/p$ this yields

$$E\sup_{t\le T}||v(t,\cdot)||^p_{L_p} \le Ne^T(||\bar f||^p_{\mathbb{L}_p(T)} + ||h||^p_{\mathbb{L}_p(T,l_2)}).$$

By using the self-similarity of the equation $dv = (v_{xx} + \bar f_x)\,dt + h^k\,dw^k_t$ (that is, by considering equations like (3.6)), for any constant $c > 0$ we get

$$E\sup_{t\le T}||v(c^2t,c\,\cdot)||^p_{L_p} \le Ne^T(||c\bar f(c^2t,c\,\cdot)||^p_{\mathbb{L}_p(T)} + ||ch(c^2t,c\,\cdot)||^p_{\mathbb{L}_p(T,l_2)})$$

with $N = N(p)$. Changing variables we obtain

$$E\sup_{t\le T}||v(t,\cdot)||^p_{L_p} \le Ne^{T/c^2}c^{p-2}(||\bar f||^p_{\mathbb{L}_p(T)} + ||h||^p_{\mathbb{L}_p(T,l_2)}).$$

For $c^2 = T$ this is even a little bit stronger than (2.7) and the lemma is proved.

Proof of Theorem 2.11. For an appropriate $\zeta \in C_0^\infty(\mathbb{R}_+)$ we have

$$E\sup_{t\le\tau}||u(t,\cdot)||^p_{H^{\gamma-1}_{p,\theta}} \le N\sum_{n=-\infty}^{\infty} e^{n\theta}E\sup_{t\le\tau}||\zeta u(t,e^n\cdot)||^p_{H^{\gamma-1}_p}. \tag{2.8}$$

Let $du = f_x\,dt + g^k\,dw^k_t$. Then

$$d(\zeta(x)u(t,e^nx)) = \zeta(x)(f_x)(t,e^nx)\,dt + \zeta(x)g^k(t,e^nx)\,dw^k_t.$$

By Lemma 2.12 for $u_n(t,x) := \zeta(x)u(t,e^nx)$, $f_n(t,x) := \zeta(x)(f_x)(t,e^nx)$, $g_n(t,x) := \zeta(x)g(t,e^nx)$, and $c = e^{-np}$ we have

$$E\sup_{t\le\tau}||u_{nx}(t,\cdot)||^p_{H^{\gamma-2}_p} \le NT^{(p-2)/2}(e^{-np}||u_{nxx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau)}$$

$$+e^{np}||f_n||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} + ||g_{nx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau,l_2)}). \tag{2.9}$$

To transform this inequality notice that all the functions $u_n(t,x)$ as functions of x have supports inside the support of ζ which is bounded. Therefore (see Remark 1.6),

$$||\zeta u(t,e^n\cdot)||_{H^{\gamma-1}_p} = ||u_n(t,\cdot)||_{H^{\gamma-1}_p} \le N||u_{nx}(t,\cdot)||_{H^{\gamma-2}_p}.$$

Furthermore, $||g_{nx}||_{H^{\gamma-2}_p(l_2)} \le ||g_n||_{H^{\gamma-1}_p(l_2)}$ and

$$\sum_{n=-\infty}^{\infty} e^{n\theta}||g_n||^p_{\mathbb{H}^{\gamma-1}_p(\tau,l_2)} \le N||g||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau,l_2)} \le N||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}.$$

Also,

$$\sum_{n=-\infty}^{\infty} e^{n(\theta+p)}||f_n||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} = \sum_{n=-\infty}^{\infty} e^{n\theta}||(M^{-1}\zeta)(MDf)(\cdot,e^n\cdot)||^p_{\mathbb{H}^{\gamma-2}_p(\tau)}$$

$$\le N||MDf||^p_{\mathbb{H}^{\gamma-2}_{p,\theta}(\tau)} \le N||f||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}(\tau)} \le N||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)},$$

$$\sum_{n=-\infty}^{\infty} e^{n(\theta-p)}||u_{nxx}||^p_{\mathbb{H}^{\gamma-2}_p(\tau)} \le \sum_{n=-\infty}^{\infty} e^{n(\theta-p)}||u_n||^p_{\mathbb{H}^{\gamma}_p(\tau)}$$

$$= \sum_{n=-\infty}^{\infty} e^{n\theta}||(M\zeta)(M^{-1}u)(\cdot,e^n\cdot)||^p_{\mathbb{H}^{\gamma}_p(\tau)}$$

$$\le N||M^{-1}u||^p_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} \le N||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}(\tau)}.$$

By combining this with (2.9) and (2.8) we get (2.4). The theorem is proved.

As always the main role is played by the spaces $\mathfrak{H}^{\gamma}_{p,\theta,0}(\tau)$ of functions with zero initial conditions. In connection with this it is worth noting that while constructing our theory we could replace

$$||u(0,\cdot)||^p_{U^{\gamma+1}_{p,\theta}} := E||M^{2/p-1}u(0,\cdot)||^p_{H^{\gamma+1-2/p}_{p,\theta}} \tag{2.10}$$

with

$$\inf\{||v_x||^p_{\mathbb{H}^{\gamma+1}_{p,\theta}} + ||M\tilde{\mathbb{D}}v||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}} + ||\tilde{\mathbb{S}}v||^p_{\mathbb{H}^{\gamma}_{p,\theta}} : u - v \in \mathfrak{H}^{\gamma+1}_{p,\theta,0}\}.$$

Such an axiomatic approach to defining a norm of $u(0,\cdot)$ yields, of course, the solvability results for the widest possible class of initial data, namely, for those which are extendible at least in some way for $t > 0$. However, in applications we often want to know how to describe "admissible" initial data by knowing only their analytic properties. A partial answer to this question is given in the following theorem, which also shows why we use the norm given by (2.10).

THEOREM 2.13. *If $0 < \theta < p$ and $\gamma = 2$ and $1 < p < \infty$, then for every u_0 satisfying $M^{2/p-1}u_0 \in H^{\gamma-2/p}_{p,\theta}$ there exists a deterministic $u \in \mathfrak{H}^{\gamma}_{p,\theta}$ such that $du = D^2u\,dt$, $u|_{t=0} = u_0$, and*

$$||u||^p_{\mathfrak{H}^{\gamma}_{p,\theta}} \le N(p,\gamma,\theta)||M^{2/p-1}u_0||^p_{H^{\gamma-2/p}_{p,\theta}}. \tag{2.11}$$

Proof. If $u_0 \in C_0^\infty(\mathbb{R}_+)$, then there is a unique function $u(t,x)$ which is bounded in $\mathbb{R}^2_+$ together with all its derivatives and which is a unique bounded solution of the heat equation $\partial u/\partial t = D^2u$, $t > 0$ in $\mathbb{R}^2_+$ with initial condition $u(0,x) = u_0(x)$ and boundary condition $u(t,0) = 0$. Observe that u is given by $u(t,\cdot) = p_t * \bar{u}_0$, where $p_t(x) = (4\pi t)^{-1/2}\exp(-|x|^2/(4t))$ and $\bar{u}_0$ is an odd extension of u_0 on $\mathbb{R}$. By the way, from this representation it follows that $u(t,x) \to 0$ exponentially fast as $x \to \infty$ and the same is true for any derivative of u.

Next, we observe that $\partial|u(t,x)|^p/\partial t = p|u|^{p-2}uD^2u$, multiply this equality by x^c, with $c := \theta + 1 - p \in (1-p, 1)$, and integrate by parts, and also use $|u(t,x)| \le N|x|$, $|u(t,x)|^{p-1}x^c \le Nx^\theta$, $|u(t,x)|^p x^{c-1} \le Nx^\theta$ for x close to zero. Finally we fix $T \in (0,\infty)$ and find that

$$\begin{aligned}\int_{\mathbb{R}_+} x^c|u(T,x)|^p\,dx - \int_{\mathbb{R}_+} x^c|u_0(x)|^p\,dx &= \int_0^T\int_{\mathbb{R}_+} px^c|u|^{p-2}uD^2u\,dxdt \\ &= -c\int_0^T\int_{\mathbb{R}_+} x^{c-1}D(|u|^p)\,dxdt - p(p-1)I = c(c-1)J - p(p-1)I,\end{aligned} \tag{2.12}$$

where

$$I := \int_0^T\int_{\mathbb{R}_+} x^c|u|^{p-2}(Du)^2\,dxdt, \quad J := \int_0^T\int_{\mathbb{R}_+} x^{c-2}|u|^p\,dxdt.$$

To estimate I from below through J, denote $v := |u|^{p/2}$ and observe that we have $|u|^{p-2}(Du)^2 = (2/p)^2(Dv)^2$ and by Minkowski's inequality

$$\int_0^\infty x^{c-2}|u|^p\,dx = \int_0^\infty x^{c-2}v^2\,dx = \int_0^\infty x^c\left(\int_0^1 v'(yx)\,dy\right)^2 dx$$

$$\le \left(\int_0^1 dy\left(\int_0^\infty x^c(v'(yx))^2\,dx\right)^{1/2}\right)^2 = \int_0^\infty x^c(v'(x))^2\,dx\left(\int_0^1 y^b\,dy\right)^2,$$

where $b = -1/2 - c/2 > -1$. By evaluating the last integral we get

$$\int_0^\infty x^{c-2}|u|^p\,dx \le p^2(1-c)^{-2}\int_0^\infty x^c|u|^{p-2}(Du)^2\,dx. \tag{2.13}$$

Hence $p(p-1)I \ge q^{-1}(1-c)^2 J$, where $1/q = 1 - 1/p$, and from (2.12) we get

$$[q^{-1}(1-c)^2 - c(c-1)]J \le \int_{\mathbb{R}_+} x^c|u_0(x)|^p\,dx = ||M^{2/p-1}u_0||^p_{L_{p,\theta}}. \tag{2.14}$$

Here $L_{p,\theta} \supset H^{2-2/p}_{p,\theta}$ with the corresponding inequality for the norms since $2-2/p > 0$. Also, one can easily check that $q^{-1}(1-c)^2 - c(c-1) > 0$ for $0 < \theta < p$ and therefore, after passing to the limit as $T \to \infty$, we obtain the following intermediate estimate:

$$\int_0^\infty ||M^{-1}u(t,\cdot)||^p_{L_{p,\theta}}\,dt \le N||M^{2/p-1}u_0||^p_{H^{2-2/p}_{p,\theta}}. \tag{2.15}$$

An attentive reader might have noticed that the above derivation of (2.13) and (2.15) falls into some trouble if $1 < p < 2$. Indeed, then we get terms containing $|u|$ to a negative power and also the absolute continuity of v is not clear. However, the following fact is true even if $1 < p < 2$:

(i) the functions $|u|^{p/2}$ and $|u|^{p-2}uu_x$ are absolutely continuous on $\mathbb{R}$;

(ii) almost everywhere on $\mathbb{R}$ ($\infty \cdot 0 := 0$)

$$(|u|^{p/2})_x = \tfrac{p}{2}|u|^{p/2-2}uu_x,$$

$$(|u|^{p-2}uu_x)_x = |u|^{p-2}uu_{xx} + (p-1)|u|^{p-2}(u_x)^2.$$

Above we have only used this fact. However, we do not prove (i) and (ii). Instead, we show how to get (2.15) for $1 < p < 2$ by using an approximation argument.

For $\varepsilon > 0$ define $G_\varepsilon(s) = (s^2+\varepsilon)^{p/2} - \varepsilon^{p/2}$. As it is easy to see, we have $|G_\varepsilon(u)| \le (1+\varepsilon^{p/2})|u|^p$ and, for $|u| \le 1$,

$$|G'_\varepsilon(u)| = p(u^2+\varepsilon)^{p/2-1}|u| \le N(\varepsilon)|u| \le N(\varepsilon)|u|^{p-1}.$$

Also $G''_\varepsilon \ge 0$. Hence, owing to $\partial G_\varepsilon(u)/\partial t = G'_\varepsilon(u)D^2u$ and introducing

$$v(t,x) := \int_0^{u(t,x)} (G''_\varepsilon(s))^{1/2}\,ds,$$

we get as above

$$\int_{\mathbb{R}_+} x^c G_\varepsilon(u(T,x))\,dx - \int_{\mathbb{R}_+} x^c G_\varepsilon(u_0(x))\,dx$$

$$= c(c-1)\int_0^T\int_{\mathbb{R}_+} x^{c-2}G_\varepsilon(u)\,dxdt - \int_0^T\int_{\mathbb{R}_+} x^c(v')^2\,dxdt$$

$$\le c(c-1)\int_0^T\int_{\mathbb{R}_+} x^{c-2}G_\varepsilon(u)\,dxdt - 4^{-1}(1-c)^2\int_0^T\int_{\mathbb{R}_+} x^{c-2}v^2\,dxdt.$$

By letting $\varepsilon \downarrow 0$, noticing that $\lim_{\varepsilon \downarrow 0} G''_\varepsilon(s) = p(p-1)|s|^{p-2}$ and $c < 1$, and using Fatou's lemma, we again arrive at (2.14) and (2.15).

Next, take a function $\zeta \in C_0^\infty(\mathbb{R}_+)$ and notice that for $u_n(t,x) := u(e^{2n}t, e^n x)$ we have

$$\frac{\partial}{\partial t}(\zeta u_n) = (\zeta u_n)_{xx} - 2(\zeta_x u_n)_x + \zeta_{xx} u_n. \tag{2.16}$$

Hence by inequalities (IV.3.1) and (IV.3.2) in [9] (also see Remark 2.3.2 in [13]) for any n we obtain

$$\int_0^\infty ||(\zeta u_n)_{xx}(t,\cdot)||^p_{H_p^{-1}}\,dt \le N||\zeta u_n(0,\cdot)||^p_{H_p^{1-2/p}}$$

$$+\int_0^\infty ||((2\zeta_x u_n)_x - \zeta_{xx}u_n)(t,\cdot)||^p_{H_p^{-1}}\,dt.$$

We make the change of variable t replacing it with $e^{2n}t$; then we multiply through the inequality by $e^{2n-np+\theta n}$ and observe that by Remark 1.6

$$||(\zeta u_n)_{xx}||_{H_p^{-1}} \ge N||(\zeta u_n)_x||_{L_p} \ge N||\zeta u_{nx}||_{L_p} - N||\zeta_x u_n||_{L_p},$$

where $N = N(\zeta, p)$. Also use the fact that

$$||(2\zeta_x u_n)_x - \zeta_{xx}u_n||_{H_p^{-1}} \le 2||(\zeta_x u_n)_x||_{H_p^{-1}} + N||\zeta_{xx}u_n||_{L_p} \le N||\eta u_n||_{L_p},$$

where $N = N(\zeta, p, \eta)$ and η is a more or less arbitrary function of class $C_0^\infty(\mathbb{R}_+)$ with support covering that of ζ.

Then we get

$$\int_0^\infty \sum_n e^{\theta n}||\zeta u_x(t,e^n\cdot)||^p_{L_p}\,dt \le N\sum_n e^{\theta n}||\xi(M^{2/p-1}u_0)(e^n\cdot)||^p_{H_p^{1-2/p}}$$

$$+N\int_0^\infty \sum_n e^{\theta n}||\eta_1 M^{-1}u(t,e^n\cdot)||^p_{L_p}\,dt,$$

where $\xi = M^{1-2/p}\zeta$ and η_1 is a function of type η. For the right choice of ζ we rewrite the last inequality as

$$\int_0^\infty ||u_x(t,\cdot)||^p_{L_{p,\theta}}\,dt \le N||M^{2/p-1}u_0||^p_{H^{1-2/p}_{p,\theta}} + N\int_0^\infty ||M^{-1}u(t,\cdot)||^p_{L_{p,\theta}}\,dt. \tag{2.17}$$

Next, we use (2.16) and inequalities (IV.3.1) and (IV.3.2) in [9] to write

$$\int_0^\infty ||(\zeta u_n)_{xx}(t,\cdot)||^p_{L_{p,\theta}}\,dt \le N||\zeta u_n(0,\cdot)||^p_{H_p^{2-2/p}}$$

$$+\int_0^\infty ||((2\zeta_x u_n)_x - \zeta_{xx}u_n)(t,\cdot)||^p_{L_{p,\theta}}\,dt.$$

If η_1 and η_2 are functions of class $C_0^\infty(\mathbb{R}_+)$ with supports covering that of ζ, then, for the same reasons as before, this inequality yields

$$\int_0^\infty ||\zeta u_{nxx}(t,\cdot)||^p_{L_{p,\theta}}\,dt \le N||\zeta u_n(0,\cdot)||^p_{H_p^{2-2/p}}$$

$$+\int_0^\infty ||\eta_1 u_{nx}(t,\cdot)||^p_{L_{p,\theta}}\,dt + \int_0^\infty ||\eta_2 u_n(t,\cdot)||^p_{L_{p,\theta}}\,dt,$$

and

$$\int_0^\infty ||Mu_{xx}(t,\cdot)||^p_{L_{p,\theta}}\,dt \le N||M^{2/p-1}u_0||^p_{H^{2-2/p}_{p,\theta}}$$

$$+N\int_0^\infty ||u_x(t,\cdot)||^p_{L_{p,\theta}}\,dt + N\int_0^\infty ||M^{-1}u(t,\cdot)||^p_{L_{p,\theta}}\,dt.$$

Together with (2.15), (2.17), and the equation $\partial u/\partial t = M^{-1}(Mu_{xx})$ the last inequality implies that $u \in \mathfrak{H}^2_{p,\theta}$ and that (2.11) holds with $\gamma = 2$.

Actually, above we have constructed a mapping $u_0 \in C_0^\infty(\mathbb{R}_+) \to u \in \mathfrak{H}^\gamma_{p,\theta}$. If we introduce an operator $\Pi : u_0 \to u$, then what is proved means that (for $\gamma = 2$)

$$||\Pi u_0||_{\mathfrak{H}^\gamma_{p,\theta}} \le N(p,\theta)||M^{2/p-1}u_0||_{H^{\gamma-2/p}_{p,\theta}} \tag{2.18}$$

if $u_0 \in C_0^\infty(\mathbb{R}_+)$. Remembering that $\mathfrak{H}^\gamma_{p,\theta}$ is a Banach space and relying on the usual continuity argument based on (2.18), we see that Π can be extended on all u_0 satisfying $M^{2/p-1}u_0 \in H^{\gamma-2/p}_{p,\theta}$ in such a way that $\partial \Pi u_0/\partial t = D^2\Pi u_0$, $\Pi u_0|_{t=0} = u_0$, and (2.18) holds. The theorem is proved.

Remark 2.14. We will see from Theorem 3.2 that Theorem 2.13 holds for any $\gamma \in \mathbb{R}$ and the solution is unique in $\mathfrak{H}^\gamma_{p,\theta}$.

In connection with this it is interesting to notice that Theorem 2.13 without weights and on $\mathbb{R}$ instead of $\mathbb{R}_+$ cannot hold for all $1 < p < 2$ if $\gamma = 1$. For instance, if $1 < p < 3/2$, then, for the solution u of the equation $du = D^2u\,dt$, $t > 0$, $x \in \mathbb{R}$, with initial condition given by the delta function, we have $u(0,\cdot) \in H_p^{1-2/p}$, but the pth power of the function u_x is not integrable over $\mathbb{R}^+ \times \mathbb{R}$.

3. SPDEs with constant coefficients on $\mathbb{R}_+$. Take a stopping time τ. On $\mathbb{R}_+$ we will be dealing with the following equation:

$$du = (au_{xx} + f_x)\,dt + (\sigma^k u_x + g^k)\,dw_t^k, \quad t \in (0,\tau), \tag{3.1}$$

where f and g^k are given $\mathcal{D}(\mathbb{R}_+)$-valued $\mathcal{P}$-measurable functions, a and σ^k are given real-valued $\mathcal{P}$-measurable functions, u is an unknown $\mathcal{D}(\mathbb{R}_+)$-valued function, and the equation is understood in the sense of distributions as follows. We say that u is a solution of (3.1) with given initial condition u_0 if for any test function $\phi \in C_0^\infty(\mathbb{R}_+)$ we have

$$\begin{aligned}(u(t,\cdot),\phi) &= (u_0,\phi)\\ &\quad + \int_0^t [a(s)(u(s,\cdot),\phi_{xx}) - (f(s,\cdot),\phi_x)]\,ds\\ &\quad + \sum_{k=1}^\infty \int_0^t [-\sigma^k(s)(u,\phi_x) + (g^k,\phi)]\,dw_t^k\end{aligned} \tag{3.2}$$

for all $t \leq \tau$ with probability one, where all integrals are assumed to have sense and the last series is also assumed to converge uniformly on each interval of time $[0, T \wedge \tau]$ in probability, where T is any finite constant.

Remark 3.1. If a function u belongs to $\mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)$, then it satisfies (3.1) with $f = (MD)^{-1}M\tilde{\mathbb{D}}u - aDu$ and $g^k = \tilde{\mathbb{S}}^k u - \sigma^k Du$. In addition (see Remark 1.5), we have $f \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$ and $g \in \mathbb{H}^{\gamma}_{p,\theta}(\tau, l_2)$. Below we show that under an additional assumption on a and σ the mapping $u \to (f, g)$ is onto.

We always assume that for some constants $K \geq \delta > 0$ and all ω, t we have

$$K \geq 2a \geq 2a - |\sigma|^2_{l_2} \geq \delta.$$

Here is the main result of this section.

THEOREM 3.2. (i) *Let* $0 < \theta < p$, $1 < p < \infty$, $\gamma \in \mathbb{R}$, $f \in \mathbb{H}^{\gamma}_{p,\theta}(\tau)$, $g \in \mathbb{H}^{\gamma}_{p,\theta}(\tau, l_2)$, *and* $u_0 \in U^{\gamma+1}_{p,\theta}$. (ii) *Assume that one of the following conditions is satisfied:*

(a) $p \geq 2$ *and* $\theta \in [p-1, p)$;

(b) $p \geq 2$ *and* $\sigma \equiv 0$;

(c) $\sigma \equiv 0$ *and* $g \equiv 0$.

Then (3.1) *with initial data* u_0 *has a unique solution in class* $\mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)$. *In addition, for this solution it holds that*

$$||u||_{\mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)} \leq N\big(||f||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} + ||g||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau,l_2)} + ||u_0||_{U^{\gamma+1}_{p,\theta}}\big), \tag{3.3}$$

where $N = N(\gamma, \theta, p, K, \delta)$. *Finally, the uniqueness holds even if we replace condition* (a) *with:* $p \geq 2$ *and* $\theta \in (0, p)$.

Remark 3.3. In a subsequent paper on equations in $\mathbb{R}^d_+$ we will show that condition (a) can be relaxed to be $p \geq 2$ and $1 \leq \theta < p$. This could be done here too if one uses interpolation with respect to θ and the result of [7], where the case $\theta = 1$ is treated. However, there is a small gap in the arguments proving (2.9) of [7], so that strictly speaking we cannot use the result of [7].

Remark 3.4. Notice that when conditions (b) or (c) are satisfied, θ may be any number in $(0, p)$.

It is also worth noting that if $\theta \geq p$ or $\theta \leq 0$, then the statement of Theorem 3.2 is false even in the case of the heat equation. This can be shown by simple examples.

The proof of this theorem is based on two lemmas, the first of which we prove in section 4.

LEMMA 3.5. *Theorem* 3.2 *holds if* $\gamma = 1$.

LEMMA 3.6. *Let assumption* (i) *of Theorem* 3.2 *be satisfied and let* $\mu \leq \gamma$. *Assume that either* $p \geq 2$ *or* $\sigma \equiv g \equiv 0$. *Let* $\theta_1 \in \mathbb{R}$ *and let* $u \in \mathfrak{H}^{\mu+1}_{p,\theta_1}(\tau)$ *be a solution of* (3.1) *with initial condition* u_0. *Assume that* $M^{-1}u \in \mathbb{H}^{\mu+1}_{p,\theta}(\tau)$. *Then* $u \in \mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)$ *and*

$$||u||_{\mathfrak{H}^{\gamma+1}_{p,\theta}(\tau)} \leq N\big(||f||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau)} + ||g||_{\mathbb{H}^{\gamma}_{p,\theta}(\tau,l_2)} + ||u_x||_{\mathbb{H}^{\mu}_{p,\theta}(\tau)} + ||u_0||_{U^{\gamma+1}_{p,\theta}}\big),$$

where $N = N(\gamma, \mu, \theta, p)$.

Proof. For simplicity of notation we will only consider the case $\tau \equiv \infty$. The reader can easily make the necessary changes for general τ.

By virtue of (3.2) we have (2.2) with $x(au_{xx} + f_x)$ instead of f and $\sigma^k u_x + g^k$ instead of g^k. Upon taking into account the assumptions on f and g and remembering Remark 1.5, we conclude that we only need to prove that

$$||u_x||^p_{\mathbb{H}^{\gamma}_{p,\theta}} \leq N(||f||^p_{\mathbb{H}^{\gamma}_{p,\theta}} + ||g||^p_{\mathbb{H}^{\gamma}_{p,\theta}(l_2)} + ||u_x||^p_{\mathbb{H}^{\mu}_{p,\theta}} + ||u_0||^p_{U^{\gamma+1}_{p,\theta}}). \tag{3.4}$$

Since $||u_x||_{\mathbb{H}^\nu_{p,\theta}} \le ||u_x||_{\mathbb{H}^\mu_{p,\theta}}$ for $\nu \le \mu$, it suffices to prove (3.4) with some $\nu \le \mu$ in place of μ. This shows that we may assume that $\gamma - \mu$ is an integer. Also we can go from μ up to γ in several steps each time getting an increase by one. Therefore, without loss of generality we may and will assume that $\gamma = \mu + 1$, so that (3.4) becomes

$$||u_x||^p_{\mathbb{H}^\gamma_{p,\theta}} \le N(||f||^p_{\mathbb{H}^\gamma_{p,\theta}} + ||g||^p_{\mathbb{H}^\gamma_{p,\theta}(l_2)} + ||u_x||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}} + ||u_0||^p_{U^{\gamma+1}_{p,\theta}}). \tag{3.5}$$

Take a function $\zeta \in C_0^\infty(\mathbb{R}_+)$ with $M\zeta$ satisfying condition (1.4). One can easily check that the functions $u_n(t,x) := u(e^{2n}t, e^n x)$ satisfy the equation

$$du_n = (a_n u_{nxx} + f_n)\,dt + (\sigma^k_n u_{nx} + g^k_n)\,dw^k_t(n), \tag{3.6}$$

where

$$a_n(t) = a(e^{2n}t), \quad \sigma^k_n(t) = \sigma^k(e^{2n}t), \quad w^k_t(n) = e^{-n} w_{e^{2n}t},$$

$$f_n(t,x) = e^{2n}(f_x)(e^{2n}t, e^n x), \quad g^k_n(t,x) = e^n g^k(e^{2n}t, e^n x).$$

Observe that for any n, the processes $w^k_t(n)$ are independent Wiener processes. From (3.6) we get

$$d(\zeta u_n) = (a_n(\zeta u_n)_{xx} + \bar f_n)\,dt + (\sigma^k_n(\zeta u_n)_x + \bar g^k_n)\,dw^k_t(n), \tag{3.7}$$

where

$$\bar f_n = \zeta f_n - 2a_n \zeta_x u_{nx} - a_n \zeta_{xx} u_n, \quad \bar g^k_n = \zeta g^k_n - \sigma^k_n \zeta_x u_n.$$

Since $M^{-1}u \in \mathbb{H}^\gamma_{p,\theta}$, it is easy to see that for any $\eta \in C_0^\infty(\mathbb{R}_+)$ we have $\eta u_n \in \mathbb{H}^\gamma_p$ and $\eta u_{nx} \in \mathbb{H}^{\gamma-1}_p$, so that $\bar f_n \in \mathbb{H}^{\gamma-1}_p$ and $\bar g_n \in \mathbb{H}^\gamma_p(l_2)$. By Theorem 2.1 of [5] or Theorem 4.10 of [8] for $p \ge 2$ (with uniqueness in $\mathcal{H}^\gamma_p(\tau)$ and existence in $\mathcal{H}^{\gamma+1}_p(\tau)$, here we use $\zeta u_n \in \mathcal{H}^\gamma_p(\tau)$), (3.7) implies that

$$||(\zeta u_n)_{xx}||^p_{\mathbb{H}^{\gamma-1}_p} \le N(||\bar f_n||^p_{\mathbb{H}^{\gamma-1}_p} + ||\bar g_n||^p_{\mathbb{H}^\gamma_p(l_2)} + E||\zeta u_0(e^n\,\cdot)||^p_{H^{\gamma+1-2/p}_p}), \tag{3.8}$$

where $u_{0n}(x) = u_0(e^n x)$. Actually, Theorem 2.1 of [5] or Theorem 4.10 of [8] treats the case $u_0 = 0$. One deals with arbitrary u_0 as in the beginning of the proof of Theorem 5.1 of [8] by just subtracting the solution of the heat equation $\partial v/\partial t = v_{xx}$ with initial condition u_0. Owing to the fact that supports of all functions ζu_n coincide with that of ζ, from (3.8) by Remark 1.6, we get

$$||\zeta u_n||^p_{\mathbb{H}^{\gamma+1}_p} \le N(||\bar f_n||^p_{\mathbb{H}^{\gamma-1}_p} + ||\bar g_n||^p_{\mathbb{H}^\gamma_p(l_2)} + E||\zeta u_0(e^n\,\cdot)||^p_{H^{\gamma+1-2/p}_p}). \tag{3.9}$$

The same conclusions are true if $1 < p < 2$ and $\sigma \equiv g \equiv 0$, which can be seen from section 9, Chapter IV of [9] or from the proof of Theorem 2.1 of [5] or Theorem 4.10 of [8], where one can take any $p \in (1,\infty)$ if $\sigma \equiv g \equiv 0$. In particular, in all cases $\zeta u_n \in \mathbb{H}^{\gamma+1}_p$ and (3.9) holds.

Now we multiply (3.9) through by $e^{(2-p+\theta)n}$ and sum up over all n. We also use

$$\sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||\zeta f_n||^p_{\mathbb{H}^{\gamma-1}_p} = \sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||e^{2n}(f_x)(e^{2n}\cdot, e^n\cdot)\zeta||^p_{\mathbb{H}^{\gamma-1}_p}$$

$$= \sum_{n=-\infty}^{\infty} e^{\theta n}||e^n(f_x)(\cdot, e^n\cdot)\zeta||^p_{\mathbb{H}^{\gamma-1}_p} = \sum_{n=-\infty}^{\infty} e^{\theta n}||(Mf_x)(\cdot, e^n\cdot)M^{-1}\zeta||^p_{\mathbb{H}^{\gamma-1}_p}$$

$$\leq N||Mf_x||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}} \leq N||f||^p_{\mathbb{H}^{\gamma}_{p,\theta}},$$

$$\sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||\zeta g_n||^p_{\mathbb{H}^{\gamma}_p(l_2)} = \sum_{n=-\infty}^{\infty} e^{\theta n}||g(\cdot, e^n\cdot)\zeta||^p_{\mathbb{H}^{\gamma}_p(l_2)} \leq N||g||^p_{\mathbb{H}^{\gamma}_{p,\theta}(l_2)},$$

$$\sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||\zeta_x u_{nx}||^p_{\mathbb{H}^{\gamma-1}_p} = \sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||e^n(u_x)(e^{2n}\cdot, e^n\cdot)\zeta_x||^p_{\mathbb{H}^{\gamma-1}_p}$$

$$= \sum_{n=-\infty}^{\infty} e^{\theta n}||(u_x)(\cdot, e^n\cdot)\zeta_x||^p_{\mathbb{H}^{\gamma-1}_p} \leq N||u_x||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}},$$

$$\sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||\zeta_{xx} u_n||^p_{\mathbb{H}^{\gamma-1}_p} = \sum_{n=-\infty}^{\infty} e^{\theta n}||(M^{-1}u)(\cdot, e^n\cdot)M\zeta_{xx}||^p_{\mathbb{H}^{\gamma-1}_p}$$

$$\leq N||M^{-1}u||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}} \leq N||u_x||^p_{\mathbb{H}^{\gamma-1}_{p,\theta}}.$$

Similarly, we estimate $\zeta_x u_n$, we notice that

$$\begin{aligned}\sum_n e^{(2-p+\theta)n} E||\zeta u_0(e^n\cdot)||^p_{H^{\gamma+1-2/p}_p} &= \sum_n e^{\theta n} E||(M^{2/p-1}u_0)(e^n\cdot)M^{1-2/p}\zeta||^p_{H^{\gamma+1-2/p}_p} \\ &\leq NE||M^{2/p-1}u_0||^p_{H^{\gamma+1-2/p}_{p,\theta}} = N||u_0||^p_{U^{\gamma+1}_{p,\theta}},\end{aligned}$$

and we get

$$\sum_{n=-\infty}^{\infty} e^{(2-p+\theta)n}||\zeta u_n||^p_{\mathbb{H}^{\gamma+1}_p} \leq I,$$

where I is the right-hand side of (3.5). Here the left-hand side equals

$$\sum_{n=-\infty}^{\infty} e^{\theta n}||(M^{-1}u)(\cdot, e^n\cdot)M\zeta||^p_{\mathbb{H}^{\gamma+1}_p} \geq N^{-1}||M^{-1}u||^p_{\mathbb{H}^{\gamma+1}_{p,\theta}} \geq N^{-1}||u_x||^p_{\mathbb{H}^{\gamma}_{p,\theta}}$$

and the lemma is proved.

Proof of Theorem 3.2. For simplicity of notation we only consider the case $\tau \equiv \infty$. Actually, as it is easy to see, the statement of existence for $\tau \equiv \infty$ implies the statement of existence for other τ, and the proof of uniqueness for general τ can be done in the same way as in the case $\tau \equiv \infty$.

Case $\gamma \geq 1$. The uniqueness follows from Lemma 3.5 and the fact that $\mathfrak{H}^2_{p,\theta,0} \supset \mathfrak{H}^{\gamma+1}_{p,\theta,0}$, which implies that the difference of two solutions belongs to $\mathfrak{H}^2_{p,\theta,0}$. The existence and estimate (3.3) follow from Lemmas 3.5 and 3.6 (applied with $\mu = 1$) and the observation that by Lemma 3.5

$$||u_x||^p_{\mathbb{H}^1_{p,\theta}} \leq N(||f||^p_{\mathbb{H}^1_{p,\theta}} + ||g||^p_{\mathbb{H}^1_{p,\theta}(l_2)} + ||u_0||^p_{U^2_{p,\theta}})$$

$$\leq N(||f||^p_{\mathbb{H}^\gamma_{p,\theta}} + ||g||^p_{\mathbb{H}^\gamma_{p,\theta}(l_2)} + ||u_0||^p_{U^{\gamma+1}_{p,\theta}}).$$

Case $\gamma < 1$. Denote by $\mathcal{R}$ the operator which maps (f, g, u_0) with $f \in \mathbb{H}^\gamma_{p,\theta}$, $g \in \mathbb{H}^\gamma_{p,\theta}(l_2)$, and $u_0 \in U^{\gamma+1}_{p,\theta}$ into the solution $u \in \mathfrak{H}^{\gamma+1}_{p,\theta}$ of (3.1) with initial data u_0. So far we know that $\mathcal{R}$ is well defined in spaces $\mathbb{H}^\gamma_{p,\theta} \times \mathbb{H}^\gamma_{p,\theta}(l_2) \times U^{\gamma+1}_{p,\theta}$ for $\gamma \geq 1$. If $\gamma < 1$, as a candidate for the solution of (3.1) we try

$$\tilde{u} = \Lambda^n_{p,\theta}\mathcal{R}(\Lambda^{-n}_{p,\theta}f, \Lambda^{-n}_{p,\theta}g, M^{1-2/p}\Lambda^{-n}_{p,\theta}M^{2/p-1}u_0),$$

where $n + \gamma \geq 1$ and (see Remark 1.3)

$$(\Lambda^{-n}_{p,\theta}f, \Lambda^{-n}_{p,\theta}g, M^{1-2/p}\Lambda^{-n}_{p,\theta}M^{2/p-1}u_0) \in \mathbb{H}^{n+\gamma}_{p,\theta} \times \mathbb{H}^{n+\gamma}_{p,\theta}(l_2) \times U^{n+\gamma+1}_{p,\theta}.$$

If the operators $\Lambda_{p,\theta}$, $M^{2/p-1}$, and D were commuting, then our candidate would be an exact solution of (3.1). Since this is not the case, we need an additional argument based on Lemma 1.14.

Take $n = 2$ and first let $1 > \gamma \geq 0$. Then by what we know in the case $\gamma \geq 1$, we have

$$v := \mathcal{R}(\Lambda^{-2}_{p,\theta}f, \Lambda^{-2}_{p,\theta}g, M^{1-2/p}\Lambda^{-2}_{p,\theta}M^{2/p-1}u_0) \in \mathfrak{H}^{\gamma+3}_{p,\theta},$$

$$dv = (av_{xx} + (\Lambda^{-2}_{p,\theta}f)_x)\,dt + (\sigma^k v_x + \Lambda^{-2}_{p,\theta}g^k)\,dw^k_t.$$

We apply $\Lambda^2_{p,\theta}$ to both parts of this equality, or in other words we substitute $(\Lambda^2_{p,\theta})^*\phi$, where $(\Lambda^2_{p,\theta})^*$ is the formal adjoint to $\Lambda^2_{p,\theta}$, in place of ϕ in (3.2). Now our candidate becomes

$$\tilde{u} = \Lambda^2_{p,\theta}v.$$

We claim that $\tilde{u}$ belongs to $\mathfrak{H}^{\gamma+1}_{p,\theta}$ and there exists

$$(\bar{f}, \bar{g}, \bar{u}_0) \in \mathbb{H}^{\gamma+1}_{p,\theta} \times \mathbb{H}^{\gamma+1}_{p,\theta}(l_2) \times U^{\gamma+2}_{p,\theta} \tag{3.10}$$

such that

$$d\tilde{u} = (a\tilde{u}_{xx} + f_x + \bar{f}_x)\,dt + (\sigma^k\tilde{u}_x + g^k + \bar{g}^k)\,dw^k_t, \tag{3.11}$$

$$\tilde{u}(0, \cdot) = u_0 + \bar{u}_0.$$

Indeed, by Remarks 2.4 and 2.8 and Lemma 1.14 we easily get that

$$\tilde{u} \in \mathfrak{H}^{\gamma+1}_{p,\theta}, \quad M^{-1}v \in \mathbb{H}^{\gamma+3}_{p,\theta}, \quad M^{-1}\tilde{u} = \Lambda^2_{p,\theta}M^{-1}v + P_1 M^{-1}v \in \mathbb{H}^{\gamma+1}_{p,\theta},$$

$$D\tilde{u} = \Lambda^2_{p,\theta}Dv + P_1 Dv \in \mathbb{H}^{\gamma}_{p,\theta}$$

and that $\tilde{u}$ satisfies (3.11) with

$$\bar{f} = 4P_2 M^{-1}v + ((2b-1)I + 2MD)\Lambda^{-2}_{p,\theta}f, \quad \bar{g}^k = \sigma^k P_1 Dv.$$

Obviously, $\bar{f}$ and $\bar{g}$ are as in (3.10). Also by Lemma 1.14 at $t = 0$,

$$M^{2/p-1}\tilde{u} = M^{2/p-1}\Lambda^2_{p,\theta}M^{1-2/p}(M^{2/p-1}v)$$

$$= \Lambda^2_{p,\theta}M^{2/p-1}v + c_1 M^{2/p-1}v + c_2 MDM^{2/p-1}v =: M^{2/p-1}u_0 + M^{2/p-1}\bar{u}_0,$$

where

$$M^{2/p-1}\bar{u}_0 \in L_p(\Omega, \mathcal{F}_0, H^{\gamma+2-2/p}_{p,\theta}).$$

This finishes the proofs of (3.10) and our claim.

Since $\gamma + 1 \geq 1$, it follows from (3.10) that the function $\bar{u} := \mathcal{R}(\bar{f}, \bar{g}, \bar{u}_0)$ is well defined, belongs to $\mathfrak{H}^{\gamma+2}_{p,\theta}$, and the function $u = \tilde{u} - \bar{u}$ is of class $\mathfrak{H}^{\gamma+1}_{p,\theta}$ and solves (3.1). For thus constructed u estimate (3.3) follows from the explicit representation and known estimates for $\mathcal{R}$, P_i, MD.

By repeating the above argument, we consider the case $0 > \gamma \geq -1$, this time using the fact that $\gamma + 1 \geq 0$ and relying upon the result for $\gamma \geq 0$. One can continue in the same way, and it only remains to prove the uniqueness of solutions in $\mathfrak{H}^{\gamma+1}_{p,\theta}$.

It suffices to consider the case $f = 0, g = 0, u_0 = 0$ (and $\gamma < 1$). In this case any solution $u \in \mathfrak{H}^{\gamma+1}_{p,\theta,0}$ also belongs to $\mathfrak{H}^{2}_{p,\theta,0}$ by Lemma 3.6 and its uniqueness follows from Lemma 3.5.

The theorem is thus proved.

Remark 3.7. In the above argument one can use $(MD)^2$ instead of $\Lambda^2_{p,\theta}$, which would make the argument shorter. We prefer $\Lambda^2_{p,\theta}$ bearing in mind a generalization to a multidimensional case.

Remark 3.8. From the above derivation of Theorem 3.2 from Lemma 3.5 it is seen that, if the assertions of Theorem 3.2 hold for some particular γ, p, θ, a, and σ satisfying the conditions of Theorem 3.2, then they hold for any $\gamma \in \mathbb{R}$ with the same p, θ, a, σ.

4. Proof of Lemma 3.5. First notice that by Theorem 2.13 for almost every ω the function $\bar{u} := \Pi u_0$ is well defined, $\bar{u} \in \mathfrak{H}^2_{p,\theta}$, $\bar{u}|_{t=0} = u_0$, $\partial\bar{u}/\partial t = \bar{f}_x$ with $\bar{f} \in \mathbb{H}^1_{p,\theta}$, and an appropriate estimate of $||\bar{u}_x||_{\mathbb{H}^1_{p,\theta}}$ and $||\bar{f}||_{\mathbb{H}^1_{p,\theta}}$ through $||u_0||_{U^2_{p,\theta}}$ holds. This implies that in the equation

$$du = (au_{xx} + (a\bar{u}_x + f - \bar{f})_x)\,dt + (\sigma^k u_x + (\sigma^k \bar{u}_x + g^k))\,dw^k_t$$

we have $a\bar{u}_x + f - \bar{f} \in \mathbb{H}^1_{p,\theta}$ and $\sigma\bar{u}_x + g \in \mathbb{H}^1_{p,\theta}(l_2)$. Also, obviously if we can solve the above equation in $\mathfrak{H}^2_{p,\theta,0}$, then by adding to the solution the function $\bar{u}$ we get a solution of (3.1) with initial data u_0. Therefore, in the proof of Lemma 3.5 without loss of generality, we may and will confine ourselves only to the case $u_0 \equiv 0$.

Furthermore, we may assume that $a \equiv 1$. Indeed, to get the result for the general case one only needs to use a random time change. Namely, let us define

$$\psi(t) = \int_0^t a(s)ds, \quad \tau(t) = \inf\{s \geq 0 : \psi(s) \geq t\},$$

$$\tilde{w}_k(t) = \int_0^{\tau(t)} \sqrt{a(s)}dw_k(s), \quad \tilde{f}(t,x) = f(\tau(t),x)/a(\tau(t)),$$

$$\tilde{\sigma}(t) = \sigma(\tau(t))/\sqrt{a(\tau(t))}, \quad \tilde{g}(t,x) = g(\tau(t),x)/\sqrt{a(\tau(t))},$$

$$\tilde{u}(t,x) = u(\tau(t),x).$$

Direct computations (see, for instance, Lemma IV.2.2 and Theorem IV.2.3 in [3]) show that $\tilde{w}_k(t)$ are independent Wiener processes and also that u is a solution of (3.1) if and only if $\tilde{u}$ is a solution of

$$d\tilde{u} = (\tilde{u}_{xx} + \tilde{f})\,dt + (\tilde{\sigma}_k\tilde{u}_x + \tilde{g}_k)\,d\tilde{w}_k(t).$$

Therefore, we easily get the desired result for general a from the result for $a \equiv 1$. Finally, obviously we may assume that $\tau \leq T$ where the constant $T < \infty$. Thus, we may and will assume that $u_0 = 0$, $a \equiv 1$, and unless stated explicitly otherwise $\tau \leq T$.

We divide the proof of the lemma in this case into the following subcases:

1. $p \geq 2$ and $\theta \in [p-1, p)$, existence;
2. $p \geq 2$ and $\theta \in (0, p)$, uniqueness;
3. $p \geq 2$ and $\sigma \equiv 0$;
4. $\sigma \equiv 0$ and $g \equiv 0$.

4.1. Case $p \geq 2$ and $\theta \in [p-1, p)$. Existence. We use the following simple lemma.

LEMMA 4.1. *Let functions f, h be defined on $\mathbb{R}_+$, be locally absolutely continuous, and satisfy*

$$\int_0^\infty |f(x)g(x)|\,dx < \infty. \tag{4.1}$$

Then

$$\int_0^\infty xf(x)g'(x)\,dx = -\int_0^\infty xf'(x)g(x)\,dx - \int_0^\infty f(x)g(x)\,dx$$

if at least one of the sides of this equality makes sense.

This fact easily follows if one integrates by parts between a, b with $0 < a < b < \infty$ and then lets $a \downarrow 0$ and $b \to \infty$ after noticing that (4.1) implies that

$$\liminf_{a\downarrow 0} |af(a)g(a)| = \liminf_{b\to\infty} |bf(b)g(b)| = 0.$$

Denote by $\mathcal{E}$ the collection of functions of the form

$$f(t,x) = \sum_{i=1}^m I_{(\tau_{i-1},\tau_i]\!]}(t)f_i(x),$$

where $f_i \in C_0^\infty(\mathbb{R}_+)$ and τ_i are stopping times, $\tau_i \le \tau_{i+1} \le \tau$. The set $\mathcal{E}$ is dense in $\mathbb{H}^1_{p,\theta}(\tau)$, which follows from a similar fact for spaces $\mathbb{H}^\gamma_p$ (see [5] or [8]) and the definition of $\mathbb{H}^\gamma_{p,\theta}(\tau)$. Also, the collection of sequences $g = (g_k)$, such that each g_k belongs to $\mathcal{E}$ and only finitely many of g_k are different from 0, is dense in $\mathbb{H}^1_{p,\theta}(\tau, l_2)$. It follows that in the proof of existence and estimate (3.3) we may assume that f and g are of this type.

Next, we use an argument from [7]. We continue $f(t,x)$ to be an even function and $g(t,x)$ to be an odd function of $x \in \mathbb{R}$. Also take an infinitely differentiable odd function $\alpha(x)$ such that $\alpha(x) = 1$ for large x, $\alpha(x) = 0$ for $|x| \le 2$ and on $\mathbb{R}$ consider the equation

$$du = (u_{xx} + f_x)\,dt + (\alpha\sigma^k u_x + g^k)\,dw^k_t. \tag{4.2}$$

The following lemma is proved in the end of this subsection.

LEMMA 4.2. *In $\mathcal{H}^0_p(\tau)$ there exists a unique solution u of* (4.2) *with zero initial condition. Moreover, $u \in \mathfrak{H}^0_{p,\theta}(\tau)$ and*

$$||u||_{\mathfrak{H}^0_{p,\theta}(\tau)} \le N||f||_{\mathbb{H}^1_{p,\theta}(\tau)} + N||g||_{\mathbb{L}_{p,\theta}(\tau,l_2)}, \tag{4.3}$$

where N is independent of τ, f, and g.

Now notice that the equation

$$du = (u_{xx} + f_x)\,dt + (\alpha_n\sigma^k u_x + g^k)\,dw^k_t, \tag{4.4}$$

where $\alpha_n(x) = \alpha(e^n x)$, also has a solution $u \in \mathfrak{H}^0_{p,\theta}(\tau)$ for which (4.3) holds with the same N. To prove this, it suffices to use scaling properties of the norms in $H^\gamma_{p,\theta}$ (see Remark 1.4) and to observe that if u is a solution of (4.2), then the function $u_n(t,x) = u(e^{2n}t, e^n x)$ satisfies (3.6) with the same f_n, g_n, and $w_t(n)$ and with $a_n = 1$ and $\sigma_n(t) = \alpha(e^n x)\sigma(e^{2n}t)$.

Denote u_n the solution of (4.4). Then u_n satisfies (4.3) and, in particular, $M^{-1}u_n$ form a bounded sequence in $\mathbb{L}_{p,\theta}(\tau)$. Denote u a weak limit of a subsequence of u_n. As in the proof of Theorem 3.11 of [8] we get that $u \in \mathfrak{H}^0_{p,\theta}(\tau)$. Then passing to the limit in (4.4) and observing that $\alpha(e^n x) \to 1$ for $x > 0$, we get that u satisfies (3.1) and estimate (4.3). It follows from Lemma 3.6 that $u \in \mathfrak{H}^2_{p,\theta}(\tau)$ and (3.3) holds with $\gamma = 1$ and $u_0 = 0$. This finishes the proof of existence.

Proof of Lemma 4.2. The existence and uniqueness of solution $u \in \mathcal{H}^1_p(\tau)$ of (4.2) is asserted in Theorem 3.2 of [5] or Theorem 5.1 of [8]. Therefore, we only need to prove that $u \in \mathfrak{H}^0_{p,\theta}(\tau)$ and that (4.3) holds.

By the definition of the norm in $\mathfrak{H}^\gamma_{p,\theta}(\tau)$ and by Remarks 2.4 and 2.7, it is sufficient to show that $M^{-1}u \in \mathbb{L}_{p,\theta}(\tau)$ and

$$||M^{-1}u||^p_{\mathbb{L}_{p,\theta}(\tau)} \le N||f||^p_{\mathbb{H}^1_{p,\theta}(\tau)} + N||g||^p_{\mathbb{L}_{p,\theta}(\tau,l_2)}. \tag{4.5}$$

Owing to our choice of f and g, from [5] or [8] we know that $u \in \mathcal{H}^\gamma_p(\tau)$ for any γ and, in particular, for almost any ω, the function $u(t,x)$ is infinitely differentiable with respect to x and all its derivatives are continuous in t. This implies that (4.2) holds pointwise (a.s.). In addition, by uniqueness the function $u(t,x)$ is odd with respect to x, so that, in particular, $u(t,0) = 0$.

Again by choice of f and g, the function u satisfies the heat equation $u_t = u_{xx}$ for $0 < x < 2$ with zero initial and zero boundary value for $x = 0$. If we set $u(t,x) = 0$

for $t<0$, then it satisfies the heat equation for all $t\leq T$ and $0<x<2$. For such functions it is well known (see, for instance, the maximum principle and Theorem 8.4.4 in [4]) that for any integer $n\geq 0$,

$$\sup_{0<x<1,t\leq T}|D^nu(t,x)|\leq N(n)\sup_{t\leq T}|u(t,2)|.$$

Therefore, for $\theta>0$,

$$E\int_0^\tau\int_0^1|u/x|^px^{\theta-1}\,dxdt\leq NE\sup_{0<x<1,t\leq T}|u_x|^p\leq NE\sup_{t\leq\tau}|u(t,2)|^p.$$

In addition, as has been mentioned above, we have $u\in\mathcal{H}^\gamma_p(\tau)$ for any γ. By embedding theorems (see [5] or [8])

$$E\sup_{[0,\tau]\times\mathbb{R}_+}|u|^p<\infty,$$

which proves that, for any $\theta>0$, we have $M^{-1}u\eta\in\mathbb{L}_{p,\theta}(\tau)$ if $\eta=\eta(x)$ is smooth and vanishes for $x\geq 1$. In the same way it is proved that for any integer $n\geq 0$ and $\theta>0$,

$$E\int_0^\tau\int_0^1|D^nu|^px^{\theta-1}\,dxdt<\infty. \tag{4.6}$$

On the other hand, $|M^{-1}u|^px^{\theta-1}\leq|u|^p$ if $x\geq 1$ and $\theta\leq p+1$. Hence, $M^{-1}u\in\mathbb{L}_{p,\theta}(\tau)$ not only for $\theta\in[p-1,p)$ but for all $\theta\in(0,p+1]$.

Next, we claim that, actually, for any $\theta\in(0,p+1]$ and $\gamma\geq 0$, we have

$$u\in\mathfrak{H}^\gamma_{p,\theta}(\tau). \tag{4.7}$$

To prove this claim, let $\zeta\in C^\infty(\mathbb{R})$ be such that $\zeta(x)=1$ for $x\leq 1/2$ and $\zeta(x)=0$ for $x\geq 1$. We want to apply Theorem 1.10 to prove that $\zeta u\in\mathfrak{H}^\gamma_{p,\theta}(\tau)$. Notice that we already know that $M^{-1}\zeta u\in\mathbb{L}_{p,\theta}(\tau)$. Also from (4.6) it follows that $M^nD^n(\zeta u)_x\in\mathbb{L}_{p,\theta}$ for any integer n. Hence the inclusion $\zeta u\in\mathfrak{H}^\gamma_{p,\theta}(\tau)$ follows indeed from Theorem 1.10.

To prove the claim it only remains to prove that $v:=(1-\zeta)u\in\mathfrak{H}^\gamma_{p,\theta}(\tau)$. Observe that $u\in\mathcal{H}^\gamma_p(\tau)$ and $v\in\mathcal{H}^\gamma_p(\tau)$ for any γ. Also, v satisfies

$$dv=(v_{xx}+\bar f)\,dt+(\alpha\sigma^kv_x+\bar g^k)\,dw^k_t, \tag{4.8}$$

where

$$\bar f=(1-\zeta)f_x+2\zeta_xu_x+\zeta_{xx}u,\quad \bar g^k=(1-\zeta)g^k+\alpha\sigma^k\zeta_xu.$$

Now, consider the following equation on $\mathbb{R}$:

$$d\tilde u=(\tilde u_{xx}-2\tilde u_x\tanh x+(2\tanh^2x-1)\tilde u+\bar f\cosh x)\,dt$$

$$+(\alpha\sigma^k\tilde u_x-\alpha\sigma^k\tilde u\tanh x+\bar g^k\cosh x)\,dw^k_t,$$

with zero initial condition. Because of compactness of supports of $\bar f$ and $\bar g$, by already cited results from [5] or [8] there is a unique solution $\tilde u$ in class $\mathcal{H}^\gamma_p(\tau)$ for any γ. Of

course, $\tilde{u}/\cosh x \in \mathcal{H}^\gamma_p(\tau)$ for any γ. In addition, one can easily check that $\tilde{u}/\cosh x$ satisfies (4.8). By the uniqueness of solutions of (4.8) in class $\mathcal{H}^\gamma_p(\tau)$, we conclude that $v = \tilde{u}/\cosh x$ and, in particular, $v\cosh x \in \mathcal{H}^\gamma_p(\tau)$ for any γ. Now the fact that $v \in \mathfrak{H}^\gamma_{p,\theta}(\tau)$ for any γ follows easily from the observation that $v = 0$ if $x \leq 1$ and $x^n/\cosh x$ is bounded.

Next we remember that (4.2) holds pointwise and we apply Itô's formula to $|u(t,x)|^p x^c$, where $c = \theta + 1 - p$. We get that, for any $x \in \mathbb{R}_+$ and $t \leq \tau$, a.s.

$$\int_0^t I(s,x)\,ds + \sum_k \int_0^t px^c|u|^{p-2}u(\alpha\sigma^k u_x - g^k)\,dw^k_s = |u(t,x)|^p x^c \geq 0, \tag{4.9}$$

where

$$I := px^{\theta-1}G(v)(xu_{xx}) + px^{\theta-1}G(v)(xf_x) + bx^{\theta-1}\sum_k |v|^{p-2}(\alpha\sigma^k u_x - g^k)^2,$$

$$b := p(p-1)/2, \quad v := u/x, \quad G(r) := |r|^{p-2}r.$$

It follows that for any $x \in \mathbb{R}_+$ there is a sequence of stopping times $\tau(n) \uparrow \tau$ localizing the stochastic integral in (4.9) so that

$$E\int_0^{\tau(n)} I(s,x)\,ds \geq 0. \tag{4.10}$$

It turns out that for almost any $x \in \mathbb{R}_+$, here one can replace $\tau(n)$ with τ and integrate with respect to x over $\mathbb{R}_+$. To prove this it suffices to prove that

$$E\int_0^\tau \int_{\mathbb{R}_+} |I(s,x)|\,dxds < \infty. \tag{4.11}$$

Observe that (4.7) for $\gamma = 2$ means that

$$M^{-1}u,\ u_x,\ Mu_{xx} \in \mathbb{L}_{p,\theta}(\tau), \tag{4.12}$$

which implies (4.11) since by Hölder's inequality

$$\begin{aligned} &E\int_0^\tau \int_0^\infty |G(v)|\,|xu_{xx}|x^{\theta-1}\,dxdt \\ &\quad = E\int_0^\tau \int_0^\infty |u(t,x)/x|^{p-1}|xu_{xx}(t,x)|x^{\theta-1}\,dxdt \\ &\quad \leq ||M^{-1}u||^{p-1}_{\mathbb{L}_{p,\theta}(\tau)}||Mu_{xx}||_{\mathbb{L}_{p,\theta}(\tau)}, \end{aligned}$$

$$E\int_0^\tau \int_0^\infty |G(v)|\,|xf_x|x^{\theta-1}\,dxdt \leq ||M^{-1}u||^{p-1}_{\mathbb{L}_{p,\theta}(\tau)}||Mf_x||_{\mathbb{L}_{p,\theta}(\tau)},$$

$$E\int_0^\tau \int_0^\infty |v|^{p-2}|u_x|^2x^{\theta-1}\,dxdt \leq ||M^{-1}u||^{p-2}_{\mathbb{L}_{p,\theta}(T)}||u_x||^2_{\mathbb{L}_{p,\theta}(\tau)},$$

$$E\int_0^\tau \int_0^\infty |v|^{p-2}|g|^2_{l_2}x^{\theta-1}\,dxdt \leq ||M^{-1}u||^{p-2}_{\mathbb{L}_{p,\theta}(\tau)}||g||^2_{\mathbb{L}_{p,\theta}(\tau,l_2)}.$$

Having thus proved (4.11), from (4.10) we conclude

$$E\int_0^\tau \int_{\mathbb{R}_+} I(s,x)\,dxds \geq 0. \tag{4.13}$$

While estimating the integral with respect to x in (4.13) we integrate by parts after noticing that (4.12) also implies that

$$E\int_0^\tau \int_0^\infty |G(u)||u_x|x^{\theta-1}\,dxdt < \infty.$$

By Lemma 4.1 for almost all $(\omega,t) \in (\!(0,\tau]\!]$ we get

$$p\int_0^\infty x^{\theta-1}G(v)(xu_{xx})\,dx = p\int_0^\infty x^c G(u)u_{xx}\,dx$$

$$= -p(p-1)\int_0^\infty |u|^{p-2}|u_x|^2x^c\,dx - c\int_0^\infty x^{c-1}(|u|^p)_x\,dx$$

$$= -p(p-1)\int_0^\infty |u|^{p-2}|u_x|^2x^c\,dx + c(c-1)\int_0^\infty |M^{-1}u|^p x^{\theta-1}\,dx.$$

Furthermore,

$$\left|\int_0^\infty x^{\theta-1}G(v)(xf_x)\,dx\right| \leq ||v||^{p-1}_{L_{p,\theta}}||Mf_x||_{L_{p,\theta}}$$
$$\leq \varepsilon||M^{-1}u||^p_{L_{p,\theta}} + N(\varepsilon,p)||f||^p_{H^1_{p,\theta}},$$

where $\varepsilon > 0$ is arbitrary. Finally, while estimating the terms in (4.13) which came from stochastic integrals we also use

$$(\alpha\sigma^k u_x - g^k)^2 \leq (1+\varepsilon)|\sigma^k|^2|u_x|^2 + (1+\varepsilon^{-1})|g^k|^2.$$

Then from (4.13) we conclude that for any $\varepsilon > 0$,

$$p(p-1)E\int_0^\tau\int_0^\infty [(1+\varepsilon)|\sigma|^2_{l_2}/2 - 1]|u|^{p-2}|u_x|^2x^c\,dxdt$$

$$+[(\theta+1-p)(\theta-p)+\varepsilon]E\int_0^\tau\int_0^\infty |M^{-1}u|^p x^{\theta-1}\,dxdt$$

$$+N(\varepsilon,p)(||f||^p_{\mathbb{H}^1_{p,\theta}(\tau)} + ||g||^p_{\mathbb{L}_{p,\theta}(\tau,l_2)})\,dxdt \geq 0. \tag{4.14}$$

Now comes the only place where we need $\theta \in [p-1,p)$. This condition implies that $(\theta+1-p)(\theta-p) \leq 0$. Also $|\sigma|^2_{l_2} \leq 2-\delta$. By using (2.13) we conclude that the first term in (4.14) is strong enough if ε is small and (4.14) implies (4.5). This brings the proof of Lemma 4.2 to an end.

4.2. Case $p \geq 2$ and $\theta \in (0, p)$. Uniqueness. Suppose that $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ is a solution of

$$du = u_{xx}\,dt + \sigma^k u_x\,dw^k_t \tag{4.15}$$

with zero initial condition. By Lemma 3.6 it follows that $u \in \mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ for all γ and also $u\zeta \in \mathcal{H}^{\gamma}_{p}(\tau)$ for all $\zeta \in C_0^\infty(\mathbb{R}_+)$. Hence we again have (4.12) and the equation is satisfied pointwise. For $\theta \in [p-1, p)$, this makes it possible to estimate the norm $||u||_{\mathfrak{H}^0_{p,\theta}(\tau)}$ using the same computations as in Lemma 4.2. Since now $f = g_k = 0$, the result is $||u||_{\mathfrak{H}^1_{p,\theta}(\tau)} = 0$.

Next notice that, for any $\theta \in (0, p)$, there exists $\theta_1 \in (p-1, p)$ such that $\theta < \theta_1 < \theta + p$. Also as above, for any γ any solution of (4.15) in $\mathfrak{H}^{\gamma}_{p,\theta}(\tau)$ with zero initial condition also belongs to $\mathfrak{H}^1_{p,\theta}(\tau)$. Hence, the following result implies the uniqueness for general $\theta \in (0, p)$.

LEMMA 4.3. *Let γ, θ_1, and p be such that the first two assertions of Theorem* 3.2 *hold for $u_0 \equiv 0$, any stopping time τ, and these γ, θ_1, and p (for instance, $\gamma = 1$, $\theta_1 \in [p-1, p)$, and $p \geq 2$). Let $q \geq p$, $\theta \neq 0$, and $\theta \neq q$ satisfy $\theta/q < \theta_1/p \leq \theta/q + 1$. Let τ be a stopping time and $u \in \mathfrak{H}^1_{q,\theta,0}(\tau)$ satisfy* (3.1) *with some $f \in \mathbb{L}_{p,\theta_1}(\tau)$ and $g \in \mathbb{L}_{p,\theta_1}(\tau, l_2)$. Then $u \in \mathfrak{H}^1_{p,\theta_1,0}(\tau)$.*

Proof. By Remark 3.8 we may assume that $\gamma = 0$. Let v be the unique solution of (3.1) in $\mathfrak{H}^1_{p,\theta,0}(\tau)$ with given f and g. To prove the lemma we prove that $u = v$.

Let κ be an infinitely differentiable function such that $\kappa(x) = 1$ for $|x| \leq 1$ and $\kappa(x) = 0$ for $|x| \geq 2$. Define $\kappa_n = \kappa(x/n)$.

First we prove that for any n,

$$u\kappa_n \in \mathfrak{H}^1_{p,\theta_1,0}(\tau). \tag{4.16}$$

To this end observe that

$$E\int_0^\tau \int_0^\infty |(u\kappa_n)_x|^p x^{\theta_1 - 1}\,dxdt \leq 2^{p-1} E\int_0^\tau \int_0^\infty |u_x\kappa_n|^p x^{\theta_1-1}\,dxdt$$

$$+2^{p-1}E\int_0^\tau \int_0^\infty |u\kappa_{nx}|^p x^{\theta_1-1}\,dxdt, \tag{4.17}$$

where by Hölder's inequality the first term on the right is less than a constant times

$$E\int_0^\tau \int_0^{2n} |u_x x^{(\theta-1)/q}|^p x^{\theta_1 - 1 - (\theta-1)p/q}\,dxdt$$

$$\leq \left(E\int_0^\tau \int_0^{2n} |u_x|^q x^{\theta-1}\,dxdt\right)^{p/q} T^{1-p/q}\left(\int_0^{2n} x^c\,dx\right)^{1-p/q},$$

with

$$c = [\theta_1 - 1 - (\theta-1)p/q]q/(q-p) = \frac{qp}{(q-p)}\left(\frac{\theta_1}{p} - \frac{\theta}{q}\right) - 1.$$

Since $c > -1$, the first term on the right in (4.17) is finite. One can similarly treat the second term after noticing that $|u\kappa_{nx}| \leq N|u/x|$ and $u/x \in \mathbb{L}_{q,\theta}(\tau)$. The same argument yields $u\kappa_n/x \in \mathbb{L}_{p,\theta_1}(\tau)$ and this proves (4.16).

Now, let $\bar{u} = u - v$. By what we have just proved, $\bar{u}\kappa_n$ belongs to $\mathfrak{H}^1_{p,\theta_1,0}(\tau)$. Also $\bar{u}\kappa_n$ satisfies the following equation similar to (3.7)

$$d(\bar{u}\kappa_n) = (a(\bar{u}\kappa_n)_{xx} + \bar{f}_{nx})\,dt + (\sigma^k(\bar{u}\kappa_n)_x + \bar{g}^k_n)\,dw^k_t,$$

where

$$\bar{f}_n(t,x) = a(t)\int_x^\infty [2\kappa_{nx}(y)\bar{u}_x(t,y) + \kappa_{nxx}(y)\bar{u}(t,y)]\,dy$$

$$= -2a\kappa_{nx}\bar{u} + (MD)^{-1}(Ma\kappa_{nxx}\bar{u}), \quad \bar{g}^k = -\sigma^k\kappa_{nx}\bar{u}.$$

Hence, by our assumptions and Remark 1.5

$$||\bar{u}\kappa_n||_{\mathfrak{H}^1_{p,\theta_1}(\tau)} \le N||\kappa_{nx}\bar{u}||_{\mathbb{L}_{p,\theta_1}(\tau)} + N||M\kappa_{nxx}\bar{u}||_{\mathbb{L}_{p,\theta_1}(\tau)}. \tag{4.18}$$

Here, for instance, ($\kappa_{nx} \le N/n$)

$$||\kappa_{nx}\bar{u}||^p_{\mathbb{L}_{p,\theta_1}(\tau)} \le Nn^{-p}E\int_0^\tau\int_n^{2n} |\bar{u}|^p x^{\theta_1-1}\,dxdt$$

$$\le NE\int_0^\tau\int_n^{2n} |v/x|^p x^{\theta_1-1}\,dxdt + Nn^{\theta_1-p-1}E\int_0^\tau\int_n^{2n} |u|^p\,dxdt.$$

The first term on the right tends to zero as $n \to \infty$ since $v/x \in \mathbb{H}^0_{p,\theta_1}(\tau)$. To prove the same for the second term use Hölder's inequality to get that it is less than

$$NT^{1-p/q}n^{\theta_1-p-p/q}\left(E\int_0^\tau\int_n^{2n} |u|^q\,dxdt\right)^{p/q}$$

$$\le Nn^c\left(E\int_0^\tau\int_n^{2n} |u|^q x^{\theta-1}\,dxdt\right)^{p/q}, \tag{4.19}$$

where $c = \theta_1 - p - p/q - (\theta-1)p/q \le 0$ by virtue of $\theta_1/p \le 1 + \theta/q$. Theorem 2.11 implies that the right-hand side of (4.19) tends to zero as $n \to \infty$.

In the same way using the fact that $|M\kappa_{nxx}| \le N/n$ we get that the second term on the right in (4.18) tends to zero as well. Thus (use Theorem 2.11)

$$E\sup_{t\le\tau}\int_0^\infty |\bar{u}(t,x)|^p x^{\theta_1-1}\,dxdt \le \liminf_{n\to\infty} E\sup_{t\le\tau}||\bar{u}(t,\cdot)\kappa_n||^p_{H^0_{p,\theta_1}}$$

$$\le N\liminf_{n\to\infty}||\bar{u}\kappa_n||^p_{\mathfrak{H}^1_{p,\theta_1}(\tau)} = 0.$$

The lemma is proved.

4.3. Case $\sigma \equiv 0$ and $p \ge 2$. Uniqueness follows directly from section 4.2. To prove existence notice that as has been emphasized in section 4.1 the only place where we used $\theta \in [p-1, p)$ is right after (4.14). But in our present situation $\sigma \equiv 0$ and from (4.14) and (2.13) we conclude that

$$[p^{-1}(p-1)(p-\theta)^2 + (\theta+1-p)(p-\theta) + \varepsilon]||M^{-1}u||^p_{\mathbb{L}_{p,\theta}(\tau)}$$

$$\le N(\varepsilon,p)(||f||^p_{\mathbb{H}^1_{p,\theta}} + ||g||^p_{\mathbb{L}_{p,\theta}(\tau,l_2)}). \tag{4.20}$$

Observe that the condition $0 < \theta < p$ is equivalent to $p^{-1}(p-1)(p-\theta)^2 + (\theta+1-p)(p-\theta) > 0$. Therefore, for ε small enough we again get (4.5). This takes care of the existence.

4.4. Case $\sigma \equiv 0$ and $g \equiv 0$. Actually, this is the case of the heat equation without any stochastic terms. In this case Lemma 3.6 is available for any $p > 1$ and as in section 4.1, to prove existence, it suffices to prove (4.5) for f as in section 4.1. This time we get (4.20) with $\varepsilon = 0$ even for $1 < p < 2$, which is proved by the same approximating argument as in the proof of Theorem 2.13 right after (2.15). Hence, we have existence.

The uniqueness is proved as in the beginning of section 4.2 observing that this time we do not need condition $\theta \in [p-1, p)$ to be satisfied and yet have (4.20).

This finishes the proof of Lemma 3.5.

REFERENCES

[1] N. V. Krylov, *A W_2^n-theory of the Dirichlet problem for SPDE in general smooth domains*, Probab. Theory Related Fields, 98 (1994), pp. 389–421.

[2] N. V. Krylov, *A generalization of the Littlewood-Paley inequality and some other results related to stochastic partial differential equations*, Ulam Quart., 2 (1994), pp. 16–26.

[3] N. V. Krylov, *Introduction to the Theory of Diffusion Processes*, AMS, Providence, RI, 1995.

[4] N. V. Krylov, *Lectures on Elliptic and Parabolic Equations in Hölder Spaces*, AMS, Providence, RI, 1996.

[5] 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.

[6] N. V. Krylov, *On SPDEs and superdiffusions*, Ann. Probab., 25 (1997), pp. 1789–1809.

[7] N. V. Krylov, *One-dimensional SPDEs with constant coefficients on the positive half axis*, in Proceedings of Steklov Mathematical Institute Seminar, Statistics and Control of Stochastic Processes, The Liptser Festschrift, Yu. Kabanov, B. Rozovskii, A. Shiryaev, eds., World Scientific, River Edge, NJ, 1997, pp. 243–251.

[8] N. V. Krylov, *An analytic approach to SPDEs*, in Stochastic Partial Differential Equations. Six Perspectives, Mathematical Surveys and Monographs, AMS, Providence, RI, to appear.

[9] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'tceva, *Linear and Quasi-Linear Parabolic Equations*, Nauka, Moscow, 1967 (in Russian); AMS, Providence, RI, 1968 (in English).

[10] S. Lapic, *On the first initial-boundary problem for SPDEs on domains with limited smoothness at the boundary*, Potential Anal., submitted.

[11] B. L. Rozovskii, *Stochastic Evolution Systems*, Kluwer, Dordrecht, 1990.

[12] E. Stein, *Harmonic Analysis: Real-Variable Methods, Orthogonality, and Oscillatory Integrals*, Princeton University Press, Princeton, NJ, 1993.

[13] H. Triebel, *Theory of Function Spaces* II, Birkhäuser-Verlag, Basel, Boston, Berlin, 1992.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 326–340

STABLE DETERMINATION OF A CRACK IN A PLANAR INHOMOGENEOUS CONDUCTOR*

GIOVANNI ALESSANDRINI† AND LUCA RONDI‡

Abstract. We prove a stability estimate for the inverse problem of cracks under essentially minimal regularity assumptions on the crack and on the background conductivity.

Key words. inverse problems, cracks, elliptic equations, quasi-conformal mappings

AMS subject classifications. 35R30, 78A30, 31A25, 30C62

PII. S0036141097325502

1. Introduction. We consider the problem of determining a crack in an electrically conducting body from current and voltage measurements at the boundary. The mathematical theory for this inverse problem was initiated by A. Friedman and M. Vogelius [F-V], who proved uniqueness theorems for a crack in a planar conductor. Stability estimates were obtained in [A2], [A3], [DV] for the case of a single crack in a homogeneous isotropic planar conductor. For an extended account on the results for this problem and for further references the reader is referred to [A-DB], where a three-dimensional theory for this problem is developed.

In this paper we prove a stability estimate for the determination of a crack in an inhomogeneous planar conductor under essentially minimal regularity assumptions on the (unknown) crack and on the (known) background conductivity.

We shall consider the conductor Ω as a simply connected bounded domain in the plane with Lipschitz boundary. The conductivity within Ω is given by a bounded and measurable tensor A which satisfies a uniform ellipticity condition. A crack σ in Ω will be a simple open curve within Ω which we shall a priori assume to be Lipschitz. Given a zero average function ψ on $\partial\Omega$, representing the prescribed current density, the electrostatic potential u in Ω will be, in the presence of the crack σ, the weak solution of the following (direct) Neumann boundary value problem:

$$\begin{cases} \operatorname{div}(A\nabla u) = 0 & \text{in } \Omega\backslash\sigma, \\ A\nabla u \cdot \nu = 0 & \text{on either side of } \sigma, \\ A\nabla u \cdot \nu = \psi & \text{on } \partial\Omega, \end{cases} \tag{1.1}$$

where ν denotes the unit normal with outward orientation when on $\partial\Omega$.

The inverse problem consists of determining the crack σ from the voltage measurements $u|_\Sigma$, Σ being a portion of $\partial\Omega$, corresponding to one or more prescribed current densities ψ.

Notice that this model corresponds to the so-called case of a perfectly insulating crack; let us stress here that our present method would enable us also to treat, with analogous results, the so-called case of perfectly conducting cracks. For the sake of brevity, we shall not discuss this case any further.

*Received by the editors August 4, 1997; accepted for publication January 27, 1998; published electronically December 11, 1998.
http://www.siam.org/journals/sima/30-2/32550.html
†Dipartimento di Scienze Matematiche, Università degli Studi di Trieste, piazzale Europa 1, 34100 Trieste, Italy (alessang@univ.trieste.it). The work of this author was partially supported by MURST and CNR.
‡SISSA-ISAS, via Beirut 2-4, 34014 Trieste, Italy (rondi@sissa.it).

As is well known since the work of A. Friedman and M. Vogelius, in order to uniquely determine σ it is necessary to perform measurements for at least two different choices ψ_1, ψ_2 of the current density ψ. We shall use here current densities ψ_1, ψ_2 analogous to those used in previous works on uniqueness [Br-V], [A-DV], [K-Se], which can be viewed as general models of a two-electrode configuration in which one electrode is kept fixed and the other is placed in two different locations. See the following section 2 for details.

Our main result (Theorem 2.1) is that the crack σ depends continuously on the boundary measurements at a rate which is of log–log type.

That is, we obtain a result which is comparable with those previously obtained for the case of a homogeneous conductor. See the concluding remarks for further details.

The present approach, however, is different from the one in [A2], [A3], [DV], which took advantage of the uniform conductivity by the use of special conformal transformations. Rather, it is closer to the approach used in [A-DV] to prove uniqueness of multiple cracks in an inhomogeneous conductor, the main novelty here being the need of stability estimates for a Cauchy problem for the elliptic equation in (1.1). In fact we shall show that the Cauchy problem for such elliptic equations has a stability character analogous to the one for the Laplace equation regardless of the smoothness of the coefficients. We shall prove this by a generalization of the classical method of harmonic measure, Theorem 4.5. We believe that this result can have some independent interest.

In section 2 we start by listing all the needed a priori assumptions and we state our main Theorem 2.1.

In section 3 we collect results based on the connections between elliptic equations in two variables, first-order Beltrami-type equations, and quasi-conformal mappings. The principal result of this section is contained in Proposition 3.7 stating Hölder continuity properties of the mappings f, f^{-1}, where f is given by $f = u + iv$, u is a solution to (1.1), and v is the associated stream function (i.e., a generalized harmonic conjugate).

Section 4 contains a treatment of a Cauchy problem and its stability properties, the main result for the rest of the paper being Proposition 4.1. Theorem 4.5 is instead a result of general type possibly useful in other contexts.

Section 5 consists of the completion of the proof of Theorem 2.1 and some concluding remarks.

2. The main theorem.

Prior information. For every $z = x + iy \in \mathbb{C}$ and for every $r > 0$ we denote with $B_r(z)$ the disk with center z and radius r. As usual, we shall identify complex numbers $z = x + iy \in \mathbb{C}$ with points $(x, y) \in \mathbb{R}^2$.

If γ is a simple curve (which could be closed) and z_0, z_1 are two points of γ, we define $\text{length}_\gamma(z_0, z_1)$ the length of the smallest arc in γ connecting z_0 to z_1.

If γ is a simple curve, r is a positive number, and z belongs to γ, we say that $\gamma \cap B_r(z)$ is a *Lipschitz graph with norm M* if there exists a system of Cartesian coordinates (x, y) with origin in z, with respect to which one has

$$\gamma \cap B_r(z) = \{(x, y) | y = \phi(x),\, x^2 + y^2 < r^2\},$$

where ϕ is a Lipschitz function on $[-r, r]$ and $\|\phi'\|_{L^\infty(-r,r)} \leq M$.

If γ is a simple open curve, r is a positive number and z is an endpoint of γ, we say that $\gamma \cap B_r(z)$ is a *half Lipschitz graph with norm M* if there exists a

system of Cartesian coordinates (x, y) with origin in z such that with respect to these coordinates one has

$$\gamma \cap B_r(z) = \{(x, y) | y = \phi(x),\ 0 \le x \le r,\ x^2 + y^2 < r^2\},$$

where ϕ is a Lipschitz function on $[0, r]$ and $\|\phi'\|_{L^\infty(0,r)} \le M$.

Let Ω be a bounded domain and $d > 0$; we denote

$$\Omega_d = \{z \in \Omega : \operatorname{dist}(z, \partial\Omega) > d\}. \tag{2.1}$$

Prior information on the domain. Let Ω be a bounded, simply connected domain in $\mathbb{R}^2$ and let its boundary $\partial\Omega$ be a simple, closed curve satisfying, for given positive constants L, δ, and M,

(2.2)(a) perimeter of $\Omega \le L$,
(2.2)(b) for every $z \in \partial\Omega$; then $\partial\Omega \cap B_\delta(z)$ is a Lipschitz graph with norm M.

Prior information on the crack. A crack σ in Ω will be a simple, open curve in Ω such that

(2.3)(a) the length of σ is less than L;
(2.3)(b) the distance of σ from $\partial\Omega$ is $\ge \delta$;
(2.3)(c) if V_1, V_2 are the endpoints of σ, then for every $i = 1, 2$ $\sigma \cap B_\delta(V_i)$ is a half Lipschitz graph with norm M; furthermore, for any $z \in \sigma \backslash (B_{\delta/2}(V_1) \cup B_{\delta/2}(V_2))$, $\sigma \cap B_{\delta/2}(z)$ is a Lipschitz graph with norm M.

Prior information on the boundary data. Let γ_0, γ_1, γ_2 be three fixed simple arcs in $\partial\Omega$, pairwise internally disjoint.

Given $\Gamma > 0$, let us fix three functions η_0, η_1, $\eta_2 \in L^2(\partial\Omega)$ such that for every $j = 0, 1, 2$,

(2.4)(a) $\eta_j \ge 0$ on $\partial\Omega$; $\operatorname{supp}(\eta_j) \subset \gamma_j$;
(2.4)(b) $\int_{\partial\Omega} \eta_j = 1$;
(2.4)(c) $\|\eta_j\|_{L^2(\partial\Omega)} \le \Gamma$.

Then we prescribe the current densities on the boundary ψ_1, ψ_2 to be given by

$$\psi_1 = \eta_0 - \eta_1, \quad \psi_2 = \eta_0 - \eta_2. \tag{2.5}$$

We have

(2.6)(a) $\int_{\partial\Omega} \psi_j = 0$ for every $j = 1, 2$;
(2.6)(b) $\|\psi_j\|_{L^2(\partial\Omega)} \le 2\Gamma$ for every $j = 1, 2$.

Moreover let us consider the following antiderivatives along $\partial\Omega$ of ψ_1, ψ_2:

$$\Psi_j(s) = \int \psi_j(s) \mathrm{d}s, \quad j = 1, 2, \tag{2.7}$$

where the indefinite integral is taken with respect to arclength on $\partial\Omega$ in the counterclockwise direction. The functions Ψ_1, Ψ_2 are defined up to an additive constant.

We remark that from the prior information on Ω, (2.2), we can find a constant M_1 depending on L, δ, and M only such that for all z_0, z_1 belonging to $\partial\Omega$ the following inequality holds:

$$\operatorname{length}_{\partial\Omega}(z_0, z_1) \le M_1 |z_0 - z_1|. \tag{2.8}$$

Hence Ψ_j verify the following property

$$|\Psi_j(z_0) - \Psi_j(z_1)| \le 2\Gamma (\operatorname{length}_{\partial\Omega}(z_0, z_1))^{1/2} \le \Gamma_1 |z_0 - z_1|^{1/2}, \tag{2.9}$$

for any z_0, z_1 belonging to the boundary of Ω, where $\Gamma_1 = 2\Gamma M_1^{1/2}$.

Prior information on the conductivity. Given $\lambda, \Lambda > 0$, let $A = A(z)$, $z \in \Omega$, be a 2×2 matrix with bounded measurable entries such that

(2.10)(a) $A(z)\xi \cdot \xi \geq \lambda |\xi|^2$ for every $\xi \in \mathbb{R}^2$ and for a.e. $z \in \Omega$;
(2.10)(b) $\|A\|_{L^\infty(\Omega)} \leq \Lambda$.

For any $i = 1, 2$, let $u_i \in W^{1,2}(\Omega \backslash \sigma)$ be the weak solution of the following Neumann-type boundary value problem:

$$\begin{cases} \operatorname{div}(A \nabla u_i) = 0 & \text{in } \Omega \backslash \sigma, \\ A \nabla u_i \cdot \nu = 0 & \text{on either side of } \sigma, \\ A \nabla u_i \cdot \nu = \psi_i & \text{on } \partial\Omega, \end{cases} \tag{2.11}$$

where ν denotes the unit normal, with the outward orientation when on $\partial\Omega$.

That is, we understand that u_i satisfies

$$\int_{\Omega \backslash \sigma} A \nabla u_i \cdot \nabla \varphi = \int_{\partial\Omega} \psi_i \varphi \quad \text{for every } \varphi \in W^{1,2}(\Omega \backslash \sigma). \tag{2.11$'$}$$

If σ' is another crack, that is, another curve satisfying conditions (2.3), we denote by u_i' the solutions to (2.11) when σ is replaced with σ'.

We denote by Σ a simple arc in $\partial\Omega$ whose length is at least δ.

The set of constants L, M, δ, Γ, λ, and Λ will be referred to as the *a priori data*.

We are now in position to state the main theorem.

THEOREM 2.1. *Under the previously stated assumptions, let $\varepsilon > 0$ be such that*

$$\max_{i=1,2} \|u_i - u_i'\|_{L^\infty(\Sigma)} \leq \varepsilon; \tag{2.12}$$

then the two cracks σ, σ' satisfy

$$\mathrm{d}_H(\sigma, \sigma') \leq \omega(\varepsilon), \tag{2.13}$$

where $\omega(\varepsilon)$ is a positive function on $(0, +\infty)$ that verifies

$$\omega(\varepsilon) \leq K(\log |\log \varepsilon|)^{-\alpha} \text{ for every } \varepsilon,\ 0 < \varepsilon < 1/e. \tag{2.14}$$

Here K and α are positive constants depending on the a priori data only.

Here d_H denotes the Hausdorff distance. We recall that the Hausdorff distance between bounded closed sets σ and σ' is given by

$$\mathrm{d}_H(\sigma, \sigma') = \max \left\{ \sup_{x \in \sigma'} \operatorname{dist}(x, \sigma), \sup_{x \in \sigma} \operatorname{dist}(x, \sigma') \right\}.$$

3. Stream functions and quasi-conformal mappings. We begin by reviewing some properties of quasi-conformal mappings which will be used in the sequel.

We shall make repeated use of the following notation for complex derivatives:

$$f_{\bar{z}} = \tfrac{1}{2}(f_x + i f_y), \quad f_z = \tfrac{1}{2}(f_x - i f_y).$$

We denote by $J = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ the counterclockwise rotation of 90° and by $(\cdot)^T$ transpose.

PROPOSITION 3.1. *Let D be a bounded simply connected domain in $\mathbb{R}^2$. Let A satisfy* (2.10). *Let $u \in W^{1,2}(D)$ be a weak solution to the equation*

$$\operatorname{div}(A \nabla u) = 0 \quad \text{in } D. \tag{3.1}$$

There exists a function $v \in W^{1,2}(D)$ which satisfies

$$\nabla v = JA\nabla u \quad \text{almost everywhere in } D. \tag{3.2}$$

Moreover, letting $f = u + iv$, we have

$$f_{\overline{z}} = \mu f_z + \nu \overline{f_z} \quad \text{almost everywhere in } D, \tag{3.3}$$

where μ and ν are bounded measurable, complex valued coefficients, satisfying

$$|\mu| + |\nu| \le k < 1 \quad \text{almost everywhere in } D, \tag{3.4}$$

where k is a constant depending on λ, Λ only.

On the other hand, if $f = u+iv$, $f \in W^{1,2}(D, \mathbb{C})$, verifies (3.3) *with coefficients μ and ν satisfying* (3.4)*, then there exists a 2×2 matrix A such that u is a weak solution of* $\operatorname{div}(A\nabla u) = 0$ *in D and A verifies* (2.10) *with constants λ, $\Lambda > 0$ depending upon k only.*

The function v appearing above is usually called the *stream function* associated with u. Notice that v is uniquely determined up to an additive constant and also that v is a weak solution to

$$\operatorname{div}(B\nabla v) = 0 \quad \text{in } D, \tag{3.5}$$

where $B = (\det A)^{-1} A^T$.

Proof. For the existence of the stream function v see [A-M, Theorem 2.1]. Then by (3.2), (3.3) follows with μ, ν given by

$$\begin{aligned} \mu &= \tfrac{a_{22}-a_{11}-i(a_{12}+a_{21})}{a_{11}a_{22}-a_{12}a_{21}+a_{11}+a_{22}+1}, \\ \nu &= \tfrac{a_{12}a_{21}-a_{11}a_{22}+1+i(a_{12}-a_{21})}{a_{11}a_{22}-a_{12}a_{21}+a_{11}+a_{22}+1}. \end{aligned} \tag{3.6}$$

From these expressions and (2.10), one obtains, through elementary although lengthy computations, (3.4).

Conversely, given the coefficients μ, ν in (3.3) satisfying (3.4) one obtains (3.1) and (3.2) with A given by

$$A = \begin{bmatrix} \frac{|1-\mu|^2-|\nu|^2}{|1+\nu|^2-|\mu|^2} & \frac{2\Im(\nu-\mu)}{|1+\nu|^2-|\mu|^2} \\ \frac{-2\Im(\mu+\nu)}{|1+\nu|^2-|\mu|^2} & \frac{|1+\mu|^2-|\nu|^2}{|1+\nu|^2-|\mu|^2} \end{bmatrix} \tag{3.7}$$

and the thesis follows. □

We recall that a *quasi-conformal map* f in an open set D is an univalent $W^{1,2}(D, \mathbb{C})$ solution of an equation of the type (3.3), (3.4).

Now we state the following representation theorem, due to L. Bers and L. Nirenberg [B-N].

THEOREM 3.2. *Let $D \subset B_1(0)$, and let $f \in W^{1,2}(D, \mathbb{C})$ verify* (3.3) *where μ, ν satisfy* (3.4).

There exists a quasi-conformal map χ from $B_1(0)$ into itself and a holomorphic function F on $\chi(D)$ such that

$$f = F \circ \chi. \tag{3.8}$$

Moreover the function χ and its inverse χ^{-1} satisfy the following conditions:

$$\begin{aligned} &|\chi(x) - \chi(y)| \le C|x-y|^\alpha \qquad \forall x, y \in B_1, \\ &|\chi^{-1}(x) - \chi^{-1}(y)| \le C|x-y|^\alpha \quad \forall x, y \in B_1, \end{aligned} \tag{3.9}$$

where C and α, $0 < \alpha < 1$, depend upon k only.

Proof. See [B-N, page 116]. □

Let us define, as in [A-M], *geometric critical points* of solutions of elliptic equations like (3.1). That is, given u as in Proposition 3.1, let v be its stream function and let χ and F, respectively, be the quasi-conformal map and the holomorphic function appearing in the representation (3.8) for $f = u + iv$.

A point $z \in \Omega$ is called a *geometric critical point* of u if $\chi(z)$ is a critical point (in the standard sense) for $\Re F$. This definition does not depend on the choice of the representation.

According to [A-M], we define the *geometric index* of u at $z \in \Omega$ as the *winding number* of F' at $\chi(z_0)$.

Remark 3.3. We wish to stress that the representation theorem, 3.2, gives us that, up to the change of coordinates χ, v can be viewed as the harmonic conjugate to u. In particular we have that, with respect to the metric in $\chi(D)$, the level lines of v are lines of steepest descent of u and vice versa. Consequently we have that, away from the discrete set of geometric critical points, u is strictly monotone on each connected component of the level lines of v, and vice versa.

The following theorem shows that, although the domain $\Omega\backslash\sigma$ is doubly connected, for the particular case of solutions to (1.1) a single valued global stream function v exists.

THEOREM 3.4. *Let u be a weak solution to* (1.1) *with $\psi \in L^2(\partial\Omega)$, $\int_{\partial\Omega} \psi = 0$.*

There exists, and it is unique up to an additive constant, a global stream function $v \in W^{1,2}(\Omega\backslash\sigma)$ related to u.

Moreover v is a weak solution of the following Dirichlet-type boundary value problem:

$$\begin{cases} \operatorname{div}(B\nabla v) = 0 & \text{in } \Omega\backslash\sigma, \\ v = \text{const} & \text{on } \sigma, \\ v = \Psi & \text{on } \partial\Omega, \\ \int_{\partial\Omega} B\nabla v \cdot \nu = 0, \end{cases} \tag{3.10}$$

where $\Psi = \int \psi ds$ on $\partial\Omega$.

Here, as above, $B = (\det(A)^{-1})A^T$. Observe that the constant value of v on σ is part of the unknowns of the problem (3.10) and that its weak formulation is to find $v \in W^{1,2}(\Omega)$ such that $v =$ constant on σ, $v = \Psi$ on $\partial\Omega$ in the sense of traces and satisfies

$$\int_\Omega B\nabla v \cdot \nabla\varphi = 0 \quad \text{for every } \varphi \in W^{1,2}(\Omega) \text{ such that } \varphi = \text{constant on } \sigma. \tag{3.10'}$$

Proof. The reader is referred to [A-DV, Proposition 2.1]. □

We note that the above theorem applies in particular to u_1, u_2 given by (2.11) and to any linear combination of such solutions. Let a, b be any two real numbers such that $a^2 + b^2 = 1$ and let us define

$$u = au_1 + bu_2, \qquad v = av_1 + bv_2, \tag{3.11}$$

$$\psi = a\psi_1 + b\psi_2, \qquad \Psi = a\Psi_1 + b\Psi_2. \tag{3.12}$$

Clearly, u is the weak solution to (1.1) and v is its stream function, solving (3.10). When σ is replaced with σ', we define u', v' in the same fashion.

Remark 3.5. Observe that, by (2.9) we have

$$|\Psi(z) - \Psi(w)| \leq 2\Gamma_1 |z - w|^{1/2} \quad \text{for every } z, w \in \partial\Omega. \tag{3.13}$$

Moreover, by (2.4), (2.5) one easily obtains that there exist points $\tilde{P}, \tilde{Q} \in \partial\Omega$ such that Ψ is monotone on the two simple curves forming $\partial\Omega \backslash \{\tilde{P}, \tilde{Q}\}$. Finally note that

$$\operatorname{osc}_{\partial\Omega} \Psi = |\Psi(\tilde{P}) - \Psi(\tilde{Q})| \geq 1/\sqrt{2}. \tag{3.14}$$

In order to distinguish the one-sided limits as $z \to \sigma$, $z \in \Omega \backslash \sigma$, it is convenient to figure out σ as a degenerate closed curve. More precisely we present Definition 3.6.

DEFINITION 3.6. *Let $\tilde{\sigma}$ be the abstract simple closed curve obtained from two copies of σ and gluing two by two the corresponding endpoints. We denote by $\tilde{\Omega}$ the compact manifold obtained by the appropriate gluing of $\overline{\Omega} \backslash \sigma$ with $\tilde{\sigma}$ and by $\tilde{d}$ the geodesic distance on $\tilde{\Omega}$.*

For any d, $p > 0$, we denote

$$\Omega_{d,p} = \{z \in \Omega | \operatorname{dist}(z, \partial\Omega) > d, \operatorname{dist}(z, \sigma) > p\}.$$

PROPOSITION 3.7. *Let $f = u + iv$, where u, v are given by* (3.11). *We have the following conditions:*

(i) *v satisfies the Hölder estimate*

$$|v(z_1) - v(z_2)| \leq C_1 |z_1 - z_2|^{\alpha_1} \quad \textit{for every } z_1, z_2 \in \overline{\Omega}. \tag{3.15}$$

(ii) *u satisfies the estimate*

$$|u(z_1) - u(z_2)| \leq C_2 (\tilde{d}(z_1, z_2))^{\alpha_1} \quad \textit{for every } z_1, z_2 \in \tilde{\Omega}. \tag{3.16}$$

(iii) *f is a quasi-conformal mapping on $\Omega \backslash \sigma$.*

(iv) *f satisfies the lower bound*

$$|f(z_1) - f(z_2)| \geq C_3(d) p^{4/\alpha_1} |z_1 - z_2|^{1/\alpha_1} \quad \textit{for every } z_1, z_2 \in \Omega_{d,p}. \tag{3.17}$$

Here C_1, C_2, $\alpha_1 > 0$ depend on the a priori data only, whereas $C_3(d) > 0$ depends on the a priori data and on d only.

Remark 3.8. It is useful to stress the difference between the estimates (3.15), (3.16). In fact, since v attains to a constant Dirichlet data on σ, it is expected that v is continuous across σ. This is not the case for u, which may have different one-sided limits on σ. This is the main motivation for the introduction of the metric $\tilde{d}$.

The proof of Proposition 3.7 will be given through several steps. At several stages we shall use the change of coordinates described below.

LEMMA 3.9. *Let Ω be a simply connected bounded open set which verifies* (2.2) *and let σ be a curve in Ω which satisfies* (2.3). *Then there exists a sense-preserving bi-Lipschitz map χ from $\Omega \backslash \sigma$ onto $B_2 \backslash \overline{B_1}$, such that the $W^{1,\infty}$ norm of χ and its inverse are dominated by constants depending on the a priori data only.*

Here and in the following we say that χ is bi-Lipschitz if it is a homeomorphism such that χ and its inverse belong to $W^{1,\infty}$.

Proof (*sketch*). First, by locally deforming $\partial\Omega$ and σ one can construct a bi-Lipschitz mapping χ_1 from Ω onto a simply connected domain Ω_1 with C^∞ boundary

such that $\sigma_1 = \chi_1(\sigma)$ is a C^∞ simple curve. Second, one can find a C^∞ diffeomorphism χ_2 from Ω_1 onto the disk $B_2(0)$ such that $\sigma_2 = \chi_2(\sigma_1)$ is the segment $\{y = 0,\ |x| \le 1/2\}$. Next, one constructs a bi-Lipschitz mapping χ_3 from the upper half disk $\overline{B_2^+(0)} = \{|z| \le 2,\ y \ge 0\}$ onto the half annulus $\overline{B_2^+(0)}\backslash B_1^+(0) = \{1 \le |z| \le 2,\ y \ge 0\}$ in such a way that $\chi_3(\sigma_2)$ is the inner half circle $\{|z| = 1,\ y \ge 0\}$ and χ_3 is the identity mapping on the rest of the boundary. Finally one can extend χ_3 as a mapping from $B_2\backslash\sigma_2$ onto $B_2\backslash\overline{B_1}$ by symmetry. One can make sure that for each χ_i, $i = 1, 2, 3$, the Jacobian and its inverse are uniformly bounded by constants depending on the a priori data only. In conclusion we pick $\chi = \chi_3 \circ \chi_2 \circ \chi_1$. □

Proof of Proposition 3.7(i). Let χ be the bi-Lipschitz map constructed in Lemma 3.9 and let us call

$$\tilde{f}(z) = f \circ \chi^{-1}, \quad z \in B_2\backslash\overline{B_1}. \tag{3.18}$$

By the $W^{1,\infty}$ bounds on χ and its inverse obtained in Lemma 3.9, χ is also quasi-conformal; hence we can find $\tilde{\mu} \in L^\infty(B_2\backslash\overline{B_1})$ such that

$$\tilde{f}_{\bar{z}} = \tilde{\mu}\tilde{f}_z \quad \text{almost everywhere in } B_2\backslash\overline{B_1}, \tag{3.19}$$

where

$$\tilde{\mu} \le \tilde{k} < 1 \tag{3.20}$$

and $\tilde{k}$ depends on the a priori data only.

Let $\tilde{v} = v \circ \chi^{-1} = \Im\tilde{f}$; then $\tilde{v}$ is a weak solution to

$$\begin{cases} \operatorname{div}(\tilde{B}\nabla\tilde{v}) = 0 & \text{in } B_2\backslash\overline{B_1}, \\ \tilde{v} = \text{const} & \text{on } \partial B_1, \\ \tilde{v} = \Psi \circ \chi^{-1} & \text{on } \partial B_2, \\ \int_{\partial B_2} \tilde{B}\nabla\tilde{v} \cdot \nu = 0, \end{cases} \tag{3.21}$$

where $\tilde{B}$ satisfies uniform ellipticity bounds of the type (2.10), with constants depending on the a priori data only.

Since the Dirichlet data in (3.21) are given as Hölder continuous traces of a $W^{1,2}(B_2\backslash\overline{B_1})$ function, by standard results of regularity up to the boundary, we obtain that $\tilde{v}$ satisfies a uniform Hölder estimate in $\overline{B_2}\backslash B_1$, with constants depending on the a priori data only.

Hence by recalling $v = \tilde{v} \circ \chi$, $\tilde{v}|_{\partial B_1} = v|_\sigma = \text{constant}$, and by the estimate

$$|\chi(z_1) - \chi(z_2)| \le C_4\tilde{d}(z_1, z_2) \quad \text{for every } z_1, z_2 \in \Omega\backslash\sigma, \tag{3.22}$$

following from Lemma 3.9, (3.15) follows. □

Proof of Proposition 3.7(ii). Let us apply the representation Theorem 3.2 to $\tilde{f}$, which gives us that, up to a quasi-conformal change of coordinates, $\tilde{u} = u \circ \chi^{-1}$ is the conjugate function to $-\tilde{v}$.

Hence by a local use of Privaloff's Theorem (see, e.g., [B-J-S, Part II, Chapter 6, Theorem 5, page 279]) we obtain that also $\tilde{u}$ satisfies a uniform Hölder estimate in $\overline{B_2}\backslash B_1$, with constants only depending on the a priori data. Hence (3.16) follows from (3.22). □

In order to proceed with the proof of (iii) of Proposition 3.7 we shall need the following two lemmas.

Let us extend $\tilde{f}$, $\tilde{\mu}$ to $B_2\backslash\overline{B_{1/2}}$ by the reflection rules

$$\begin{cases} \tilde{f}(z) = \overline{\tilde{f}(1/\overline{z})} + 2ci, \\ \tilde{\mu}(z) = \overline{\tilde{\mu}(1/\overline{z})}, \end{cases} \qquad z \in B_2\backslash\overline{B_{1/2}}, \tag{3.23}$$

where $c = \tilde{v}|_{\partial B_1}$.

We obtain that $\tilde{f} \in W^{1,2}(B_2\backslash\overline{B_{1/2}}, \mathbb{C})$ and satisfies (3.19) on all of $B_2\backslash\overline{B_{1/2}}$ where $|\tilde{\mu}| \le \tilde{k} < 1$ obviously holds throughout. Note that (3.23) imply that $\tilde{u}$, $\tilde{v}$ satisfy the reflection rules

$$\begin{cases} \tilde{u}(z) = \tilde{u}(1/\overline{z}), \\ \tilde{v}(z) = 2c - \tilde{v}(1/\overline{z}), \end{cases} \qquad z \in B_2\backslash\overline{B_{1/2}}, \tag{3.24}$$

and according to Proposition 3.1 are solutions to uniformly elliptic equations in all of $B_2\backslash\overline{B_{1/2}}$.

LEMMA 3.10. *$\tilde{u}$ has exactly two geometric critical points $\tilde{P}_1$, $\tilde{P}_2$ of index one in $B_2\backslash\overline{B_{1/2}}$. $\tilde{P}_1$, $\tilde{P}_2$ belong to ∂B_1 and they are distinct.*

Remark. It may be useful to stress that $\tilde{P}_1$, $\tilde{P}_2$ are also the unique geometric critical points of $\tilde{v}$ in $B_2\backslash\overline{B_{1/2}}$.

Proof. This statement is proven in [A-DV, Proposition 3.2] except from the fact that $\tilde{P}_1$, $\tilde{P}_2$ are distinct. This can be obtained by the following contradiction argument, if we had $\tilde{P}_1 = \tilde{P}_2$ then, on $\partial B_1\backslash\{\tilde{P}_1\}$, $\tilde{v} \equiv$ *constant* and hence $\tilde{u}$ should be strictly monotone along such a simple curve, thus contradicting its continuity at $\tilde{P}_1$. □

Let us denote $m = \min_{\partial\Omega} \Psi$, $M = \max_{\partial\Omega} \Psi$, and $c = v|_\sigma$. Observe that by the use of the maximum principle in (3.10′) one obtains $m < c < M$.

LEMMA 3.11. *For any $t \in (m, M)$, $t \neq c$, the level line $\{z \in \Omega\backslash\sigma |\, v(z) = t\}$ is composed by a simple curve γ_t joining the two connected components of the level set $\{z \in \partial\Omega |\, \Psi(z) = t\}$.*

The level line $\{z \in \Omega\backslash\sigma |\, v(z) = c\}$ is composed of two simple curves γ_c^1, γ_c^2 each joining σ with one of the two connected components of $\{z \in \partial\Omega |\, \Psi(z) = c\}$, respectively. Moreover the limit points of γ_c^1, γ_c^2 on σ are given by two single points P_1, P_2 which are distinct as elements of $\tilde{\sigma}$.

Proof. By the continuity (3.15) of v we have that for every $t \in (m, M)$ the limit points of $\{z \in \Omega\backslash\sigma |\, v(z) = t\}$ on $\partial\Omega \cup \tilde{\sigma}$ all belong to $\{z \in \partial\Omega |\, \Psi(z) = t\}$ if $t \neq c$ and to $\{z \in \partial\Omega |\, \Psi(z) = c\} \cup \tilde{\sigma}$ if $t = c$.

Let $t \neq c$ and let $z_0 \in \Omega\backslash\sigma$ be such that $v(z_0) = t$. By Lemma 3.10 we have that $v = \tilde{v} \circ \chi$ has no geometric critical points in $\Omega\backslash\sigma$. Therefore, by the maximum principle, the connected component γ_t of $\{v = t\}$ containing z_0 is a simple curve having endpoints on $\partial\Omega$. Again, by the maximum principle, we obtain that $v \neq t$ outside of γ_t and hence $\{v = t\} = \gamma_t$. By the same reasoning, we may find two distinct arcs γ_c^1, γ_c^2 in $\Omega\backslash\sigma$ on which $v = c$, each joining σ to the two distinct components of $\{z \in \partial\Omega |\, \Psi(z) = c\}$. Such curves disconnect $\Omega\backslash\sigma$, and hence, by the maximum principle, they exhaust the level set $\{v = c\}$. Concerning the limit points of $\{v = c\}$ on σ, these coincide with the preimages through χ of the geometric critical points $\tilde{P}_1$, $\tilde{P}_2$ of $\tilde{v}$, and the thesis follows. □

Proof of Proposition 3.7(iii). It suffices to prove that f is univalent. We use the notation introduced in Lemma 3.11. Let $\tilde{\sigma}_1$, $\tilde{\sigma}_2$ the abstract simple curves forming $\tilde{\sigma}\backslash\{P_1, P_2\}$. Using the representation $u = \tilde{u} \circ \chi$ and the absence of geometric critical points for $\tilde{u}$ in $B_2\backslash(\overline{B_{1/2}} \cup \{\tilde{P}_1, \tilde{P}_2\})$ we have that u is strictly increasing on each of

the curves $\gamma_c^1 \cup \gamma_c^2 \cup \tilde{\sigma}_i$, $i = 1, 2$. Analogously, when $t \in (m, M)$, $t \neq c$, u is strictly increasing on γ_t. Therefore for any $\zeta = s + it \in f(\Omega \backslash \sigma)$ there exists a unique $z \in \Omega \backslash \sigma$ such that $v(z) = t$, $u(z) = s$. □

Proof of Proposition 3.7(iv). With the aid of Theorem 3.2 and of a suitable conformal mapping, we obtain that there exist $R > 1$ depending on the a priori data only and a quasiconformal mapping χ_1 from $B_2 \backslash \overline{B_{1/2}}$ onto $B_R \backslash \overline{B_1}$ and a holomorphic function F on $B_R \backslash \overline{B_1}$ such that

$$\tilde{f} = F \circ \chi_1;$$

moreover, χ_1, χ_1^{-1} satisfy uniform Hölder estimates with constants depending on the a priori data only.

Let U, V be the real and imaginary part of F, respectively.

We remark that, in view of Lemma 3.10, F has exactly two critical points, which are distinct and have multiplicity one, in $B_R \backslash \overline{B_1}$. We denote such points $\zeta_1 = \chi_1(\tilde{P}_1)$, $\zeta_2 = \chi_1(\tilde{P}_2)$. Let us denote $D = B_R \backslash \overline{B_1}$ and $D_d = B_{R-d} \backslash \overline{B_{1+d}}$, $d > 0$. We claim the following lower bound on $|F'|$, whose proof is deferred to the end of this section.

Claim. There exists a positive constant C_5 depending on the a priori data and on d only such that the following estimate holds

$$|F'(z)| \geq C_5 |z - \zeta_1||z - \zeta_2| \quad \text{for any } z \in D_d. \tag{3.25}$$

Let us now recall that $F = f \circ \chi^{-1} \circ \chi_1^{-1}$ and let us fix d, $p > 0$. Denote by γ the image through χ_1 of ∂B_1, that is, $\gamma = (\chi_1 \circ \chi)(\tilde{\sigma})$. Let $\alpha_2 > 0$ be a uniform Hölder exponent for $\chi_1 \circ \chi$ and its inverse. We recall that α_2 depends on the a priori data only.

For any $z \in \Omega_{d,p}$ we have

$$\begin{aligned} &\operatorname{dist}(\chi_1 \circ \chi(z), \partial B_R) \geq C_6 d^{1/\alpha_2}, \\ &\operatorname{dist}(\chi_1 \circ \chi(z), \gamma) \geq C_6 p^{1/\alpha_2}, \end{aligned}$$

where C_6 depends on the a priori data only.

We remark that $F(\gamma) = f(\sigma)$ is a horizontal segment l.

So using (3.25) we can show that the image through f of $\Omega_{d,p}$ is contained in a doubly connected open set $D_1 \subset f(\Omega \backslash \sigma)$ whose boundary is constituted by two curves γ_1 and γ_2. The outer one, γ_1, is a Jordan curve such that for any z_0, $z_1 \in \gamma_1$ the following estimate holds

$$\operatorname{length}_{\gamma_1}(z_0, z_1) \leq C_7 |z_0 - z_1|.$$

On the other hand, γ_2 is the set of points whose distance from the segment l is equal to $C_8 p^{3/\alpha_2}$.

Furthermore on D_1 we can find the following estimate:

$$|(F^{-1})'|(z) \leq C_9 p^{-3/\alpha_2} \quad \text{for any } z \in D_1.$$

Then, evaluating the geodetic distance on D_1, we have that for any z, $w \in D_1$ it holds that

$$|F^{-1}(z) - F^{-1}(w)| \leq C_{10} p^{-4/\alpha_2} |z - w|; \tag{3.26}$$

hence for any z, $w \in \Omega_{d,p}$ we have

$$|f(z) - f(w)| \geq C_{11} p^{4/\alpha_2} |z - w|^{1/\alpha_2}. \tag{3.27}$$

The constants C_7–C_{11} depend on d and on the a priori data only. So (3.17) follows. □

Proof of the Claim. We adapt arguments used in [A1, Theorem 1.3]. First, we notice that F is Hölder continuous in D. Hence $|F|$ can be bounded on D by a constant C_{12}, C_{12} depending on the a priori data only, and in view of (3.14) there exists d_1 small enough such that for any $0 < d \le d_1$ the oscillation of V on ∂D_d is greater than $1/2\sqrt{2}$.

Without loss of generality we can restrict our attention to the case $0 < d \le d_1$; then, by using estimates on Cauchy's integrals, we have

$$|F'| \le 2C_{12}/d \qquad \forall z \in D_{d/2}, \tag{3.28}$$

$$|F''| \le 8C_{12}/d^2 \qquad \forall z \in D_{d/2}. \tag{3.29}$$

We denote $\phi = \log \frac{|F'|}{|z-\zeta_1||z-\zeta_2|}$; this is a harmonic function in D. Let $M = \sup_{D_{d/2}} \phi$; then we apply the Harnack inequality to $M - \phi$ and obtain

$$\sup_{D_d}(M-\phi) \le c \inf_{D_d}(M-\phi),$$

where c depends on d and on R only. This, in turn, implies that

$$\inf_{D_d} \phi \ge M - c(M - \sup_{D_d} \phi). \tag{3.30}$$

Notice that we have

$$1/2\sqrt{2} \le \operatorname{osc}_{\partial D_d} V \le C_{13} \max_{D_d} |F'| \le C_{14} \max_{D_d} \exp \phi,$$

and hence $M \ge C_{15} > 0$. Using (3.29), possibly choosing a smaller value for the constant d_1, we can find an upper bound on $M - \sup_{D_d} \phi$. Hence we can find a constant C_{16}, depending on the a priori data and on d only, such that $\inf_{D_d} \phi \ge C_{16}$ and the claim follows. □

4. Stability for a Cauchy problem. Let u be given by (3.11) and let u' be given accordingly when σ is replaced with σ'. Let v and v' be the stream functions of u and u', respectively; we choose to normalize v, v' in such a way that they have the same Dirichlet data Ψ on $\partial\Omega$.

Let us denote $\Phi = W + iZ = u - u' + i(v - v') : \Omega \backslash (\sigma \cup \sigma') \mapsto \mathbb{C}$.

We have that Z is identically zero on $\partial\Omega$ and $|W| \le \sqrt{2}\varepsilon$ on Σ. We remember that, by Proposition 3.7(i), (ii), there exists a constant K_1 depending on the a priori data only such that

$$|\Phi(z)| \le K_1 \quad \text{for any } z \in \Omega \backslash (\sigma \cup \sigma'). \tag{4.1}$$

Furthermore by (3.15) the function Z is Hölder continuous on $\overline{\Omega}$ with constants depending on the a priori data only.

Φ satisfies the Cauchy problem

$$\begin{cases} \Phi_{\overline{z}} = \mu \Phi_z + \nu \overline{\Phi_z} & \text{in } \Omega \backslash (\sigma \cup \sigma'), \\ |\Phi| \le \sqrt{2}\varepsilon & \text{on } \Sigma, \\ \Im \Phi = 0 & \text{on } \partial\Omega, \end{cases} \tag{4.2}$$

where $|\mu| + |\nu| \le k < 1$.

We want to estimate $|Z|$ on $\overline{\Omega}$ in terms of ε.

PROPOSITION 4.1. *Under the previous assumptions we have*

$$|Z(z)| \leq \eta(\varepsilon) \quad \text{for any } z \in \overline{\Omega}, \tag{4.3}$$

where η is a positive function defined on $(0, +\infty)$ that verifies

$$\eta(\varepsilon) \leq K_2(\log|\log\varepsilon|)^{-\beta_1} \quad \text{for every } \varepsilon,\ 0 < \varepsilon < 1/e. \tag{4.4}$$

Here K_2 and β_1 are positive constants depending on the a priori data only.

Let us recall some notions from potential theory; see for instance the book by J. Heinonen, T. Kilpeläinen, and O. Martio, [H-K-Ma].

Let D be a bounded open set. Let $A \in L^\infty(D)$ be a 2×2 matrix which satisfies (2.10).

We denote by $\mathcal{L}_A$ the differential operator

$$\mathcal{L}_A u = -\mathrm{div}(A\nabla u). \tag{4.5}$$

DEFINITION 4.2. *A function $u : D \mapsto \mathbb{R} \cup \{+\infty\}$ is called* $\mathcal{L}_A$-superharmonic *in D if*

(i) *u is lower semicontinuous;*

(ii) *$u \not\equiv +\infty$ in any connected component of D;*

(iii) *for any open set $D_1 \subset\subset D$ and any $h \in C(\overline{D_1})$, such that $\mathcal{L}_A h = 0$ in the weak sense in D_1, if $u \geq h$ on ∂D_1 then $u \geq h$ in D_1.*

A function u is $\mathcal{L}_A$-subharmonic *in D if $-u$ is $\mathcal{L}_A$-superharmonic in D.*

DEFINITION 4.3. *Let E be a subset of ∂D and let χ_E be its characteristic function. We define the* $\mathcal{L}_A$-harmonic measure *of E with respect to D as the* upper Perron solution *with respect to χ_E; that is,*

$$\omega(z) = \omega(E, D, \mathcal{L}_A; z) = \inf\{u(z) |\, u \in \mathcal{U}_E\} \quad \text{for any } z \in D,$$

where $\mathcal{U}_E$ is the class of the $\mathcal{L}_A$-superharmonic functions u in D such that $u \geq 0$ and $\liminf_{x\to y} u(x) \geq \chi_E(y)$ for any $y \in \partial D$.

LEMMA 4.4. *Let D be a bounded domain. Let $f \in W^{1,2}(D, \mathbb{C})$ satisfy* (3.3), (3.4). *There exists a 2×2 matrix $A_1 \in L^\infty(D)$ satisfying* (2.10) *with constants λ, Λ depending on k only such that $\phi = \log|f|$ is $\mathcal{L}_{A_1}$-subharmonic.*

Proof. Let z be a point in D such that $f(z) \neq 0$. Locally, on a neighborhood of z, we can define the function $\phi_1 = \log f$ where log is any possible determination of the logarithm in the complex plane.

In this neighborhood ϕ_1 verifies the equation

$$(\phi_1)_{\overline{z}} = \mu(\phi_1)_z + \nu_1\overline{(\phi_1)_z}, \tag{4.6}$$

where $\nu_1 = \nu\overline{f}/f$ and hence $|\mu| + |\nu_1| \leq k < 1$.

Then we consider the matrix A_1 corresponding to μ and ν_1, as in (3.7). By Proposition 3.1 the function $\phi = \log|f| = \Re\log f$ locally verifies

$$\mathrm{div}(A_1\nabla\phi) = 0 \tag{4.7}$$

in the weak sense.

We remark that we can define $\phi = \log|f|$ globally as a $W^{1,2}_{loc}(D_1)$ function, where $D_1 = \{z \in D |\, f(z) \neq 0\}$; hence using a partition of unity it is easy to show that (4.7) holds weakly in D_1.

Clearly the set $\{z \in D |\, f(z) = 0\}$ consists of isolated points and ϕ goes uniformly to $-\infty$ as z converges to an element of such a set.

Using this remark and the maximum principle, we can prove in an elementary way that $\phi = \log|f|$ is $\mathcal{L}_{A_1}$-subharmonic. □

By the use of suitable $\mathcal{L}_{A_1}$-harmonic measure we obtain a Hölder stability estimate in the interior for Cauchy problems like (4.2), as follows.

THEOREM 4.5. *Let D be bounded domain and E a subset of ∂D. Let f satisfy* (3.3), (3.4).

If $C = \sup|f|$ on D and we have that, given $\varepsilon > 0$,

$$\limsup_{x \to y} |f(x)| \leq \varepsilon \tag{4.8}$$

for any $y \in E$, then for any $z \in D$ the following estimate holds

$$|f(z)| \leq C^{1-\omega(z)} \varepsilon^{\omega(z)}, \tag{4.9}$$

where $\omega = \omega(E, D, \mathcal{L}_{A_1})$ is the $\mathcal{L}_{A_1}$-harmonic measure of E with respect to D and the matrix A_1 is defined as in the thesis of the Lemma 4.4.

Proof. We can assume, without loss of generality, that $0 < \varepsilon < C$. Consider the function $\phi = \log|f|$, by the fact that Lemma 4.4 ϕ is $\mathcal{L}_{A_1}$-subharmonic. Let $\omega = \omega(E, D, \mathcal{L}_{A_1})$ be the $\mathcal{L}_{A_1}$-harmonic measure of E with respect to D.

Let us denote $\phi_2 = \frac{\phi - \log(C)}{\log(\varepsilon) - \log(C)}$. It is easy to see that ϕ_2 belongs to the upper class $\mathcal{U}_E$. Hence for any $z \in D$ we have $\omega(z) \leq \phi_2(z)$ and so

$$\phi(z) \leq \log(\varepsilon)(\omega(z)) + \log(C)(1 - \omega(z)). \tag{4.10}$$

And this clearly implies the thesis. □

Remark. Observe that in view of Proposition 3.1 the above Theorem 4.5 could be restated in terms of a Cauchy problem for an elliptic equation like (3.1).

Proof of Proposition 4.1 (*Sketch*). The proof of this proposition can be obtained along the same lines as in the proof of Theorem 3.1 in [A2], once Theorem 4.5 is available.

First consider curves γ, with the first endpoint on Σ, whose h-neighborhoods γ_h are contained in $\Omega \backslash (\sigma \cup \sigma')$.

Then we apply Theorem 4.5 inside such domains γ_h, and we consider a point $z \in \gamma_h$ and $\omega = \omega(\Sigma \cap \partial\gamma_h, \gamma_h; \mathcal{L}_{A_1})$ as in Theorem 4.5. We obtain, recalling (4.1), (4.2),

$$|\Phi(z)| \leq K_1^{1-\omega(z)} \varepsilon^{\omega(z)}.$$

We find a positive lower bound on $\omega(z)$ by a repeated use of the Harnack inequality; then through Hölder continuity of Z in $\overline{\Omega}$ we can evaluate an upper bound for $|Z|$ on $\overline{\gamma_h}$. Finally we use the maximum principle together with the fact that v and v' are constant on σ, σ', respectively, to obtain the desired bound for $|Z|$ on $\overline{\Omega}$. □

5. Proof of the main Theorem 2.1. The proof of Theorem 2.1 will be completed by combining Proposition 4.1 with the following result. □

PROPOSITION 5.1. *Let all the assumptions of Theorem* 2.1 *be satisfied with the exception of* (2.12). *Let v_i be the stream functions related to u_i and let v_i' be those related to u_i'. If we have*

$$\max_{i=1,2} \|v_i - v_i'\|_{L^\infty(\Omega)} \leq \eta, \tag{5.1}$$

then the two cracks σ, σ' satisfy

$$\mathrm{d}_H(\sigma,\sigma') \le K_3\eta^{\beta_2}, \tag{5.2}$$

where K_3, β_2, $K_3 > 0$, $0 < \beta_2 < 1$, only depend on the a priori data.

Proof. Up to reversing the role of σ and σ' we may fix $z_0 \in \sigma'\backslash\sigma$ in such a way that $p = \mathrm{dist}(z_0,\sigma) = d_H(\sigma,\sigma') > 0$.

There exists a positive constant $K_4 > 1$ only depending on the a priori data such that

(5.3)(a) $B_{p/K_4}(z_0) \subset \Omega_{\delta/2}\backslash\sigma$;
(5.3)(b) there exists a point $z_1 \in \sigma'$ such that $|z_1 - z_0| = p/2K_4$.

Hence we can determine two real numbers a, b such that $a^2 + b^2 = 1$ and

$$au_1(z_o) + bu_2(z_0) = au_1(z_1) + bu_2(z_1) \tag{5.4}$$

holds true.

So we define u and v as in (3.11) and it turns out that

$$u(z_o) = u(z_1). \tag{5.5}$$

Recall that u solves (1.1) and v is its stream function. Let, as usual, $f = u + iv$.

Then by (iv) of Proposition 3.7 there exists a constant K_5, depending on the a priori data only, such that

$$p^{5/\alpha_1} \le K_5|f(z_0) - f(z_1)|. \tag{5.6}$$

Note that, by (5.5), $|f(z_0) - f(z_1)| = |v(z_0) - v(z_1)|$. We have that z_0 and z_1 belong to σ'; hence $v'(z_0) = v'(z_1)$.

So we have

$$|f(z_0) - f(z_1)| \le 2\eta. \tag{5.7}$$

Consequently

$$p \le K_6\eta^{\alpha_1/5}, \tag{5.8}$$

where K_6 and α_1 only depend on the a priori data. ∎

Concluding remarks. Let us recall that, for the case of uniform background conductivity, a log-log–type stability like the present one was proven in [A2]. Subsequently, in [A3], it was shown that the stability could be improved to a log-type estimate. A $C^{2,\alpha}$ a priori bound on σ was assumed. It can be verified that the approach in [A3] could be used with minor adaptations also in the present case, at the cost of assuming a somewhat stronger a priori assumption on the crack. For instance, an analysis of this sort has been developed in [R] where it was assumed a $C^{1,\alpha}$ bound on σ and a Lipschitz bound on A. Let us stress here that in view of Theorem 4.5 in this paper any regularity assumption on A can be dropped.

Let us recall here also the examples in [A4] for the so-called inverse problem of corrosion detection, which is different, but strictly allied, to the crack problem. Such examples show that logarithmic stability is best possible for that problem and they strongly suggest that this is the case also for the crack problem.

From another point of view, we notice that the Lipschitz regularity assumptions on σ and on $\partial\Omega$ could be further relaxed. In fact we could cast our analysis within

the theory of *quasicircles* (see [P, Chapter 5] and [L]) and prescribe that σ and $\partial\Omega$ satisfy the so-called arc condition. This ensures that quasi-conformal mappings in $\Omega\backslash\tilde{\sigma}$ are Hölder continuous up to the boundary, thus permitting us to derive statements analogous to Proposition 3.7 and, consequently, to Theorem 2.1. However, we have preferred to confine ourselves within the Lipschitz class which, we believe, is sufficiently wide and manageable from the applications point of view.

REFERENCES

[A1] G. Alessandrini, *An identification problem for an elliptic equation in two variables,* Ann. Mat. Pura Appl., 145 (1986), pp. 265–296.

[A2] G. Alessandrini, *Stable determination of a crack from boundary measurements,* Proc. Roy. Soc. Edinburgh Sect. A, 123 (1993), pp. 497–516.

[A3] G. Alessandrini, *Stability for the crack determination problem,* in Inverse Problems in Mathematical Physics, L. Päivärinta and E. Somersalo, eds., Springer-Verlag, Berlin, Heidelberg, 1993, pp. 1–8.

[A4] G. Alessandrini, *Examples of instability in inverse boundary-value problems,* Inverse Problems, 13 (1997), pp. 887–897.

[A-DB] G. Alessandrini and E. Di Benedetto, *Determining 2-dimensional cracks in 3-dimensional bodies*: *Uniqueness and stability,* Indiana Univ. Math. J., 46 (1997), pp. 1–82.

[A-DV] G. Alessandrini and A. Diaz Valenzuela, *Unique determination of multiple cracks by two measurements,* SIAM J. Control Optim., 34 (1996), pp. 913–921.

[A-M] G. Alessandrini and R. Magnanini, *Elliptic equations in divergence form, geometric critical points of solutions, and Stekloff eigenfunctions,* SIAM J. Math. Anal., 25 (1994), pp. 1259–1268.

[B-J-S] L. Bers, F. John, and M. Schechter, *Partial Differential Equations,* Interscience, New York, London, Sydney, 1964.

[B-N] L. Bers and L. Nirenberg, *On a representation theorem for linear elliptic systems with discontinuous coefficients and its applications,* in Convegno Internazionale sulle Equazioni Lineari alle Derivate Parziali, Trieste, Cremonese, Roma, 1955.

[Br-V] K. Bryan and M. Vogelius, *A uniqueness result concerning the identification of a collection of cracks from finitely many boundary measurements,* SIAM J. Math. Anal., 23 (1992), pp. 950–958.

[DV] A. Diaz Valenzuela, *Unicità e stabilità per il problema inverso del crack perfettamente isolante,* thesis, Università degli Studi di Trieste, 1993.

[F-V] A. Friedman and M. Vogelius, *Determining cracks by boundary measurements,* Indiana Univ. Math. J., 38 (1989), pp. 527–556.

[H-K-Ma] J. Heinonen, T. Kilpeläinen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations,* Clarendon Press, Oxford, New York, Tokyo, 1993.

[K-Se] H. Kim and J. K. Seo, *Unique determination of a finite number of collection of cracks from two boundary measurements,* SIAM J. Math. Anal., 27 (1996), pp. 1336–1340.

[L] O. Lehto, *Univalent Functions and Teichmüller Spaces,* Springer-Verlag, New York, Berlin, Heidelberg, 1987.

[P] Ch. Pommerenke, *Boundary Behaviour of Conformal Maps,* Springer-Verlag, Berlin, Heidelberg, 1992.

[R] L. Rondi, *Stabilità per il problema inverso dei crack in un corpo non omogeneo,* thesis, Università degli Studi di Trieste, 1996.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 341–359

THE BREAKDOWN OF SUPERCONDUCTIVITY DUE TO STRONG FIELDS FOR THE GINZBURG–LANDAU MODEL*

T. GIORGI† AND D. PHILLIPS‡

Abstract. We study the behavior of a superconducting material subjected to a constant applied magnetic field, $\mathbf{H}_a = h\mathbf{e}$ with $|\mathbf{e}| = 1$, using the Ginzburg–Landau theory. We analytically show the existence of a critical field $\overline{h}$, for which when $h > \overline{h}$, the normal states are the only solutions to the Ginzburg–Landau equations. We estimate $\overline{h}$. As $\kappa \downarrow 0$ we derive $\overline{h} = O(1)$, while as $\kappa \to \infty$ we obtain $\overline{h} = O(\kappa)$.

Key words. superconductivity, Ginzburg–Landau equations, upper critical fields, normal state

AMS subject classifications. 35J60, 35J65, 35Q40

PII. S0036141097323163

1. Introduction.

If a superconducting body is subjected to a sufficiently strong applied magnetic field, its ability to act as a superconductor breaks down and only the normally conducting (resistive) state is observed. In this paper we consider superconductivity as modeled by the Ginzburg–Landau theory and establish this type of phenomena. Here, superconductivity is characterized in terms of a complex valued order parameter, ψ (where $|\psi|^2$ represents the density of superconducting electron pairs), and a vector field $\mathbf{A}$—the magnetic potential.

Consider a superconducting body given by a bounded domain $\boldsymbol{\mathcal{D}} \subset \mathbb{R}^n$, where $n = 2$ or 3 and $\partial\boldsymbol{\mathcal{D}}$ is of class $C^{2,\alpha}$ for some $0 < \alpha < 1$. Assume the body has constant permeability normalized equal to one and that the exterior consists of a second material with constant permeability $\mu_e > 0$. Define the permeability density as follows:

$$\begin{aligned} \mu(\mathbf{x}) &= 1 \quad \text{for } \mathbf{x} \in \boldsymbol{\mathcal{D}} \\ &= \mu_e \quad \text{for } \mathbf{x} \in \mathbb{R}^n \setminus \overline{\boldsymbol{\mathcal{D}}}. \end{aligned}$$

A magnetic field is applied to all space in the form $\mathbf{H}_a = h\mathbf{e}$, where h is a positive constant and $\mathbf{e} \in \mathbb{R}^3$ is a fixed unit vector. The presence of $\boldsymbol{\mathcal{D}}$ produces an induced magnetic field, $\frac{1}{\mu}\operatorname{curl}\mathbf{A}$ in $\mathbb{R}^3$, and a supercurrent density $\mathbf{j} := \frac{-i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) - \mathbf{A}|\psi|^2$ in $\boldsymbol{\mathcal{D}}$. Here $\kappa > 0$ is the Ginzburg–Landau constant determined from the superconducting material and the superscript $*$ denotes complex conjugation. According to this theory, the pair $(\psi, \mathbf{A})$ is an equilibrium state for the Gibbs free

*Received by the editors June 20, 1997; accepted for publication (in revised form) February 18, 1998; published electronically, January 5, 1999. This research was sponsored in part by NSF Grant #9622305-DMS and also supported by the Department of Energy Grant #DE-FG02-90ER45427 through the Midwest Superconductivity Consortium.

http://www.siam.org/journals/sima/30-2/32316.html

†Department of Mathematics and Statistics, McMaster University, Hamilton, ON L8S 4K1, Canada (giorgi@icarus.math.mcmaster.ca).

‡1395 Department of Mathematics, Purdue University, West Lafayette, IN 47907-1395 (phillips@math.purdue.edu).

energy

$$G(\psi,\mathbf{A}) := \int_{\boldsymbol{\mathcal{D}}} \left(\left| \frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi \right|^2 + \frac{1}{2}(1-|\psi|^2)^2 \right) d\mathbf{x} + \int_{\mathbb{R}^n} \mu \left| \frac{1}{\mu}\operatorname{curl}\mathbf{A} - h\mathbf{e} \right|^2 d\mathbf{x} + \frac{\gamma}{\kappa}\int_{\partial\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{s} \tag{1.1}$$

(see [5], [14]). The constant $\gamma \geq 0$ reflects the retarding effect of the material in the exterior domain on the density $|\psi|^2$ at $\partial\boldsymbol{\mathcal{D}}$; γ is taken to be zero if $\mathbb{R}^n \setminus \overline{\boldsymbol{\mathcal{D}}}$ is a vacuum and large if the exterior is a magnetic material. Thus, we consider pairs $(\psi, \mathbf{A})$ such that

$$\psi \in H^1(\boldsymbol{\mathcal{D}};\mathbb{C}) \equiv \mathcal{H}^1(\boldsymbol{\mathcal{D}}), \quad \mathbf{A} \in H^1_{\mathrm{loc}}(\mathbb{R}^n;\mathbb{R}^n),$$

which are weak solutions to

$$\begin{cases} \left(\dfrac{i}{\kappa}\nabla + \mathbf{A}\right)^2 \psi - \psi + |\psi|^2\psi = 0 & \text{in } \boldsymbol{\mathcal{D}}, \\ \operatorname{curl}\left(\dfrac{1}{\mu}\operatorname{curl}\mathbf{A}\right) + \left(\dfrac{i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \mathbf{A}|\psi|^2\right)\chi_{\boldsymbol{\mathcal{D}}} = 0 & \text{in } \mathbb{R}^n, \\ \mathbf{n}\cdot\left(\dfrac{i}{\kappa}\nabla + \mathbf{A}\right)\psi = -i\gamma\psi & \text{on } \partial\boldsymbol{\mathcal{D}}, \\ \left(\dfrac{1}{\mu}\operatorname{curl}\mathbf{A} - h\mathbf{e}\right) \in L^2(\mathbb{R}^n;\mathbb{R}^3). \end{cases} \tag{1.2}$$

Here $\mathbf{n}$ is the outward normal to $\boldsymbol{\mathcal{D}}$ at $\partial\boldsymbol{\mathcal{D}}$ and $\chi_{\boldsymbol{\mathcal{D}}}$ is the characteristic function for $\boldsymbol{\mathcal{D}}$.

A principal feature of the energy (1.1) and the solutions to (1.2) is that they are invariant under the gauge transformation

$$(\psi,\mathbf{A}) \to (\psi',\mathbf{A}'),$$

where

$$\psi' = \psi e^{i\kappa\eta}, \quad \mathbf{A}' = \mathbf{A} + \nabla\eta$$

for an arbitrary real valued function $\eta \in H^2_{\mathrm{loc}}(\mathbb{R}^n)$. Moreover, the intrinsic quantities for a solution are preserved under this transformation: its density $|\psi'|^2 = |\psi|^2$, magnetic field $\frac{1}{\mu}\operatorname{curl}\mathbf{A}' = \frac{1}{\mu}\operatorname{curl}\mathbf{A}$, current $\mathbf{j}' = \mathbf{j}$, and the modulus of the derivative $|(\frac{i}{\kappa}\nabla + \mathbf{A}')\psi'| = |(\frac{i}{\kappa}\nabla + \mathbf{A})\psi|$.

A solution is in the *normal phase* if $\psi \equiv 0$ in $\boldsymbol{\mathcal{D}}$. This is written as $(\psi,\mathbf{A}) = (0, h\mathbf{a}_N)$, where $\mathbf{a}_N$ satisfies

$$\begin{aligned} &\operatorname{curl}\left(\frac{1}{\mu}\operatorname{curl}\mathbf{a}_N\right) = 0 \quad \text{in } \mathbb{R}^n, \\ &\left(\frac{1}{\mu}\operatorname{curl}\mathbf{a}_N - \mathbf{e}\right) \in L^2(\mathbb{R}^n;\mathbb{R}^3). \end{aligned} \tag{1.3}$$

Such a solution is called a *normal state*. It is uniquely determined by μ and $\boldsymbol{\mathcal{D}}$ up to a gauge transformation; that is, (1.3) uniquely determines $\operatorname{curl}\mathbf{a}_N$.

Let κ be fixed. We denote $\overline{h}$ as the *upper critical field* for the body

$$\overline{h} := \inf\{h' : \text{ normal states are the only solutions to (1.2) for all } h > h'\}.$$

For the case that the body is a bounded domain $\boldsymbol{\mathcal{D}} \subset \mathbb{R}^3$, we prove the following statement:

Let $\boldsymbol{\mathcal{D}} \subset \mathbb{R}^3$. Given κ, μ_e, and γ we have $\overline{h} = \overline{h}(\kappa, \mu_e, \gamma, \boldsymbol{\mathcal{D}}) < \infty$ (see Theorem 3.12).

We show the normal induction is continuous on $\overline{\boldsymbol{\mathcal{D}}}$. In the case that it does not vanish on $\overline{\boldsymbol{\mathcal{D}}}$, we can estimate $\overline{h}$.

If $\operatorname{curl} \mathbf{a}_N \neq \mathbf{0}$ in $\overline{\boldsymbol{\mathcal{D}}} \subset \mathbb{R}^3$, then there are constants m, $\phi \geq 0$, depending on μ_e and $\boldsymbol{\mathcal{D}}$ so that

$$\overline{h}(\kappa, \mu_e, \gamma, \boldsymbol{\mathcal{D}}) \leq \max\left(\frac{m}{\kappa}, \phi\kappa\right) \tag{1.4}$$

(see Theorem 3.9).

In the classic case where $\mu_e = 1$, it follows that $\operatorname{curl} \mathbf{a}_N \equiv \mathbf{e}$ and as such (1.4) applies.

If $\mu \equiv 1$, then there are constants m and ϕ such that $\overline{h} \leq \max(\frac{m}{\kappa}, \phi\kappa)$ (see Corollary 3.10).

We also consider the case of a cylindrical domain of the form $\boldsymbol{\mathcal{D}} \times \mathbb{R}$ where the cross section, $\boldsymbol{\mathcal{D}}$, is a bounded domain in $\mathbb{R}^2$ with a $C^{2,\alpha}$ boundary and the applied field $\mathbf{H}_a = h\mathbf{e} = h\mathbf{e}_3$ is perpendicular to the cross section. From symmetry the problem reduces to one in two dimensions. We consider $\psi(x, y)$ for $(x, y) \in \boldsymbol{\mathcal{D}}$ and $\mathbf{A} = (A_1(x, y), A_2(x, y))$ for $(x, y) \in \mathbb{R}^2$. The functional (1.1) then represents the Gibbs free energy per unit length for the cylinder. We prove the following theorem.

Let $\boldsymbol{\mathcal{D}} \times \mathbb{R}$ be a cylindrical body in a parallel applied field $h\mathbf{e}_3$. Given κ, μ_e, and γ there is a finite upper critical field $\overline{h}$, so that if $h > \overline{h}$ then the only solution to (1.2) *with $n = 2$ is normal. Moreover, there is a constant $\phi(\mu_e, \boldsymbol{\mathcal{D}})$ so that $\overline{h}(\kappa, \mu_e, \gamma, \boldsymbol{\mathcal{D}}) \leq \max(\frac{1}{\kappa}, \phi\kappa)$* (see Theorem 2.9).

Finally, we consider the case of small κ. We prove the following result.

Let $n = 2$ with $\mu_e > 0$ or $n = 3$ with $\mu_e = 1$. Then $\overline{h} = O(1)$ as $\kappa \downarrow 0$ (see Theorem 4.1).

It is of interest to compare these results with conjectures made by physicists. For κ fixed, de Gennes and St. James have studied the local problem of determining the smallest value of h for which all normal states are stable for $h' \geq h$. The infimum, denoted as h_{c_3}, is the value for which it is possible to have a family of superconducting solutions bifurcate away from the normal state. In [13] they discussed the case of an infinite slab $-d < x < d$, $-\infty < y, z < \infty$ in $\mathbb{R}^3$. The symmetry of the domain reduced the linear analysis to a one-dimensional problem. They gave an ansatz for determining h_{c_3} and predicted $\lim_{\kappa \to \infty} h_{c_3}/\kappa = c_0$ for some constant $1 < c_0 < 2$. This can be compared with our estimates for $\overline{h}$ from Theorems 2.9 and 3.9. We have $h_{c_3} \leq \overline{h} = O(\kappa)$ as $\kappa \to \infty$. For small κ, physicists have predicted that $\overline{h} = O(1)$ as $\kappa \to 0$ for a slab of finite thickness $-d < x < d$, $-\infty < y, z < \infty$ and that $\overline{h} = O(\kappa^{-\frac{1}{2}})$ as $\kappa \to 0$ for the infinitely thick slab $-\infty < x < 0$, $-\infty < y, z < \infty$ (see [4], [8], and [12]). Our estimate from Theorem 4.1 gives the result $\overline{h} = O(1)$ as $\kappa \to 0$ for our domain $\boldsymbol{\mathcal{D}}$.

To conclude, we comment on past analytic work. In [2] and [3], Bolley and Bolley and Helffer made the ansatz for the slab rigorous and proved asymptotic estimates for h_{c_3}. In [4] they obtained partial results for estimating an upper critical field for the slab. They considered a particular family of one-dimensional functions. For each fixed κ, they showed there is a finite upper critical field when considering only solutions in this family. In [1] Bauman, Phillips, and Tang estimated h_{c_3} for the case of a circular

cylinder, $B_r \times \mathbb{R}$. This estimate is relevant here as it plays a central role in our analysis of $\overline{h}$ for general domains.

In section 2 we consider cylindrical domains and establish Theorem 2.9. In section 3 we extend these ideas so as to treat bounded domains in $\mathbb{R}^3$. In section 4 we estimate $\overline{h}$ for small κ.

2. Superconductivity within an infinitely long cylinder in a parallel field. Let $(\psi, \mathbf{A})$ be a weak solution to (1.2) with $n = 2$ and $\boldsymbol{\mathcal{D}} \subset \mathbb{R}^2$. Recall that $\mathbf{H}_a = h\mathbf{e}_3$ is perpendicular to the cross section. We first examine the magnetic induction, $\operatorname{curl} \mathbf{A}$, in $\boldsymbol{\mathcal{D}}^c$.

LEMMA 2.1. *Let* $(\psi, \mathbf{A})$ *satisfy* (1.2). *Then* $\operatorname{curl} \mathbf{A}$ *is constant in each component of* $\overline{\boldsymbol{\mathcal{D}}}^c$. *Moreover,* $\operatorname{curl} \mathbf{A} = \mu_e h \mathbf{e}_3$ *in the unbounded component.*

Proof. From (1.2) we see that $\operatorname{curl}(\operatorname{curl} \mathbf{A}) = \mathbf{0}$ in each component of $\overline{\boldsymbol{\mathcal{D}}}^c$. Since

$$\operatorname{curl}(\operatorname{curl} \mathbf{A}) = (D_y(D_x A_2 - D_y A_1), -D_x(D_x A_2 - D_y A_1), 0),$$

we have that $\operatorname{curl} \mathbf{A} = (D_x A_2 - D_y A_1)\mathbf{e}_3$ is constant in each of these components. The last assertion follows from the fourth equation in (1.2). ∎

We now determine $\operatorname{curl} \mathbf{a}_N$.

LEMMA 2.2. *A normal state exists. Moreover, any normal state* $(0, h\mathbf{a}_N)$ *satisfies* $\operatorname{curl} \mathbf{a}_N = \mu \mathbf{e}_3$.

Proof. Consider $w = \Gamma_2 * (\mu - \mu_e)$, where $\Gamma_2(\mathbf{x}) = \frac{1}{2\pi} \ln(|\mathbf{x}|)$, $\mathbf{x} = (x, y)$. The function $\mu - \mu_e$ has bounded support. As a result w is well defined with $w \in H^2_{\text{loc}}(\mathbb{R}^2)$ and $\Delta w = (\mu - \mu_e)$ in $\mathbb{R}^2$. Set $\mathbf{b}_N = (-w_y, w_x) + \frac{\mu_e}{2}(-y, x)$. Then $\operatorname{curl} \mathbf{b}_N = \mu \mathbf{e}_3$ and we see $\mathbf{b}_N$ is a weak solution to (1.3), that is,

$$\int_{\mathbb{R}^2} \frac{1}{\mu} \operatorname{curl} \mathbf{b}_N \operatorname{curl} \varphi \, d\mathbf{x} = 0, \text{ for all } \varphi \in H^1(\mathbb{R}^2; \mathbb{R}^2)$$

such that φ has bounded support. Thus, a normal state exists.

Suppose $\mathbf{a}_N$ is another weak solution. Taking the difference of the equations for $\mathbf{b}_N$ and $\mathbf{a}_N$ we get

$$\int_{\mathbb{R}^2} \frac{1}{\mu} \operatorname{curl}(\mathbf{b}_N - \mathbf{a}_N) \cdot \operatorname{curl} \varphi \, d\mathbf{x} = 0. \tag{2.1}$$

Let $\boldsymbol{\mathcal{E}}$ be the unbounded component of $\overline{\boldsymbol{\mathcal{D}}}^c$. From Lemma 2.1 we have $\operatorname{curl} \mathbf{b}_N = \operatorname{curl} \mathbf{a}_N$ outside of the bounded set $\boldsymbol{\mathcal{E}}^c$. As a result we can take φ such that $\varphi = \mathbf{b}_N - \mathbf{a}_N$ in a neighborhood of $\boldsymbol{\mathcal{E}}^c$ in (2.1). Whence, $\operatorname{curl} \mathbf{b}_N \equiv \operatorname{curl} \mathbf{a}_N$. ∎

We can now show a weak solution has a gauge equivalent representative that satisfies a Sobolev estimate.

LEMMA 2.3. *Let* $(\zeta, \mathbf{B})$ *and* $(0, h\mathbf{a}_N)$ *be weak solutions to* (1.2). *Then there is a weak solution* $(\psi, \mathbf{A})$ *that is gauge equivalent to* $(\zeta, \mathbf{B})$ *such that*

$$\int_{\boldsymbol{\mathcal{D}}} |\mathbf{A} - h\mathbf{a}_N|^2 d\mathbf{x} \le C_0 \int_{\mathbb{R}^2} |\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)|^2 d\mathbf{x} \tag{2.2}$$

where C_0 *depends only on* $\boldsymbol{\mathcal{D}}$.

Proof. Set $\operatorname{curl}(\mathbf{B} - h\mathbf{a}_N) = f\mathbf{e}_3$. From Lemmas 2.1 and 2.2, we have that the support of f is contained in the bounded set $\boldsymbol{\mathcal{E}}^c$ and $f \in L^2(\mathbb{R}^2)$. Set $w = \Gamma_2 * f$; then standard estimates on the Newtonian potential give $w \in H^2_{\text{loc}}(\mathbb{R}^2)$, $\nabla w = \nabla \Gamma_2 * f$, $\|\nabla w\|_{L^2(\boldsymbol{\mathcal{E}}^c)} \le C_0(\boldsymbol{\mathcal{D}}) \|f\|_{L^2(\boldsymbol{\mathcal{E}}^c)}$, and $\Delta w = f$ (see [9]). Thus, setting $\widetilde{\mathbf{A}} = (-w_y, w_x)$

we have $\widetilde{\mathbf{A}} \in H^1_{\mathrm{loc}}(\mathbb{R}^2;\mathbb{R}^2)$ and $\operatorname{curl}\widetilde{\mathbf{A}} = \Delta w \mathbf{e}_3 = \operatorname{curl}(\mathbf{B} - h\mathbf{a}_N)$. Let $\mathbf{A} = \widetilde{\mathbf{A}} + h\mathbf{a}_N$. Then $\operatorname{curl}(\mathbf{B} - \mathbf{A}) = \mathbf{0}$. Hence, $\mathbf{A} = \mathbf{B} + \nabla\eta$ for some $\eta \in H^2_{\mathrm{loc}}(\mathbb{R}^2)$ and

$$\int_{\boldsymbol{\mathcal{D}}} |\mathbf{A} - h\mathbf{a}_N|^2 d\mathbf{x} \le \int_{\boldsymbol{\mathcal{E}}^c} |\nabla w|^2 d\mathbf{x} \le C_0 \int_{\boldsymbol{\mathcal{E}}^c} |f|^2 d\mathbf{x} = C_0 \int_{\mathbb{R}^2} |\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)|^2 d\mathbf{x}. \qquad \square$$

We need the following property for weak solutions.

PROPOSITION 2.4. (see [7]). *Let $(\psi, \mathbf{A})$ be a weak solution to* (1.2); *then $|\psi| \le 1$ almost everywhere in $\boldsymbol{\mathcal{D}}$.*

Next we write the weak formulation of (1.2):

$$\begin{aligned} &\int_{\boldsymbol{\mathcal{D}}} \left(\frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi\right) \cdot \left(\frac{i}{\kappa}\nabla\varphi + \mathbf{A}\varphi\right)^* d\mathbf{x} + \int_{\boldsymbol{\mathcal{D}}} (|\psi|^2 - 1)\psi\varphi^* d\mathbf{x} \\ &\qquad = -\frac{\gamma}{\kappa}\int_{\partial\boldsymbol{\mathcal{D}}} \psi\varphi^* d\mathbf{s} \qquad \text{for any } \varphi \in \mathcal{H}^1(\boldsymbol{\mathcal{D}}), \\ &\int_{\mathbb{R}^2} \frac{1}{\mu}\operatorname{curl}\mathbf{A}\cdot\operatorname{curl}\mathbf{B}\, d\mathbf{x} + \int_{\boldsymbol{\mathcal{D}}} \left[\frac{i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \mathbf{A}|\psi|^2\right]\cdot\mathbf{B}\, d\mathbf{x} = 0 \end{aligned} \tag{2.3}$$

for any $\mathbf{B} \in H^1(\mathbb{R}^2;\mathbb{R}^2)$ with bounded support. Considering $\frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi$ for $(\psi, \mathbf{A}) \in \mathcal{H}^1(\boldsymbol{\mathcal{D}}) \times H^1_{\mathrm{loc}}(\mathbb{R}^2;\mathbb{R}^2)$, we have

$$\begin{aligned} \Re e\left[\left(\frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi\right)\psi^*\right] &= \frac{i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \mathbf{A}|\psi|^2, \\ \Im m\left[\left(\frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi\right)\psi^*\right] &= \frac{1}{2\kappa}\nabla|\psi|^2. \end{aligned}$$

Thus,

$$\frac{i}{\kappa}\nabla\psi + \mathbf{A}\psi = \left\{\left[\frac{i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \mathbf{A}|\psi|^2\right]|\psi|^{-1} + i\left[\frac{1}{\kappa}\nabla|\psi|\right]\right\}\frac{\psi}{|\psi|} \tag{2.4}$$

for almost every $\mathbf{x}$ such that $\psi \neq 0$. Moreover, since $\nabla\psi = 0$ almost everywhere on the set $\{\psi = 0\}$, it is consistent to define the term in braces equal to zero on this set. We conclude that (2.4) holds almost everywhere.

LEMMA 2.5. *Let $(\psi, \mathbf{A})$ and $(0, h\mathbf{a}_N)$ be weak solutions satisfying* (2.2). *Then there is a constant $C_1 = C_1(\boldsymbol{\mathcal{D}}, \mu_e)$ such that*

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \kappa h\mathbf{a}_N)\psi|^2 d\mathbf{x} \le C_1\kappa^2 \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x}. \tag{2.5}$$

Proof. Let $\varphi = \psi$ in (2.3.1). Using (2.4), Proposition 2.4, and $\gamma \ge 0$, we obtain

$$\begin{aligned} &\int_{\boldsymbol{\mathcal{D}}} \left(\left|\frac{1}{\kappa}\nabla|\psi|\right|^2 + \left|\left[\frac{i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \mathbf{A}|\psi|^2\right]|\psi|^{-1}\right|^2\right) d\mathbf{x} \\ &\qquad = \int_{\boldsymbol{\mathcal{D}}} \left|\left(\frac{i}{\kappa}\nabla + \mathbf{A}\right)\psi\right|^2 d\mathbf{x} \le \int_{\boldsymbol{\mathcal{D}}} (1 - |\psi|^2)|\psi|^2 d\mathbf{x} \le \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x}. \end{aligned} \tag{2.6}$$

This inequality is also valid for $n = 3$. Consider the second equation in (1.2) for the solutions $(\psi, \mathbf{A})$ and $(0, h\mathbf{a}_N)$. Taking the difference of their respective weak equations we have

$$\int_{\mathbb{R}^2} \frac{1}{\mu}\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)\cdot\operatorname{curl}\mathbf{B}\, d\mathbf{x} = -\int_{\boldsymbol{\mathcal{D}}} \left[\frac{i}{2\kappa}(\psi^*\nabla\psi - \psi\nabla\psi^*) + \mathbf{A}|\psi|^2\right]|\psi|^{-1}\cdot|\psi|\mathbf{B}\, d\mathbf{x}.$$

Using (2.6) and the Cauchy–Schwarz inequality, we see

$$\int_{\mathbb{R}^2} \frac{1}{\mu} \operatorname{curl}(\mathbf{A} - h\mathbf{a}_N) \cdot \operatorname{curl} \mathbf{B}\, d\mathbf{x} \leq \varepsilon^{-1} \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} + \varepsilon \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 |\mathbf{B}|^2 d\mathbf{x}$$

for any $\varepsilon > 0$. Let $\mathbf{B}$ be such that $\mathbf{B} = \mathbf{A} - h\mathbf{a}_N$ in $\boldsymbol{\mathcal{E}}^c$ where $\boldsymbol{\mathcal{E}}$ is the unbounded component of $\boldsymbol{\mathcal{D}}^c$. Then since $\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N) = \mathbf{0}$ in $\boldsymbol{\mathcal{E}}$ and $|\psi| \leq 1$ we derive

$$\int_{\mathbb{R}^2} \frac{1}{\mu} |\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)|^2 d\mathbf{x} \leq \varepsilon^{-1} \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} + \varepsilon \int_{\boldsymbol{\mathcal{D}}} |A - h\mathbf{a}_N|^2 d\mathbf{x}.$$

Combining this inequality with (2.2), we see we can take ε sufficiently small so to have

$$\int_{\boldsymbol{\mathcal{D}}} |\mathbf{A} - h\mathbf{a}_N|^2 d\mathbf{x} \leq M \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} \tag{2.7}$$

for some constant $M = M(\operatorname{diam} \boldsymbol{\mathcal{D}}, \mu_e)$.

Next we write

$$\left(\frac{i}{\kappa}\nabla + \mathbf{A}\right)\psi = \left(\frac{i}{\kappa}\nabla + h\mathbf{a}_N\right)\psi + (\mathbf{A} - h\mathbf{a}_N)\psi. \tag{2.8}$$

We will use the elementary inequality

$$\frac{1}{2}|\mathbf{c}|^2 - |\mathbf{b}|^2 \leq |\mathbf{c} + \mathbf{b}|^2 \quad \text{for } \mathbf{c}, \mathbf{b} \in \mathbb{C}. \tag{2.9}$$

Let $(\frac{i}{\kappa}\nabla + \mathbf{A})\psi = \mathbf{b}$ and $-(\frac{i}{\kappa}\nabla + h\mathbf{a}_N)\psi = \mathbf{c}$. Then using (2.6) and (2.8) we derive

$$\frac{1}{2}\int_{\boldsymbol{\mathcal{D}}} \left|\left(\frac{i}{\kappa}\nabla + h\mathbf{a}_N\right)\psi\right|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} + \int_{\boldsymbol{\mathcal{D}}} |\mathbf{A} - h\mathbf{a}_N|^2 |\psi|^2 d\mathbf{x}.$$

Since $|\psi| \leq 1$, we can apply (2.7) to obtain

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \kappa h\mathbf{a}_N)\psi|^2 d\mathbf{x} \leq 2(1 + M)\kappa^2 \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x}.$$

We set $C_1 = 2(1 + M)$ and the lemma is proved. □

We see that if a superconducting state (i.e., a solution with $\psi \not\equiv 0$) exists; then (2.5) implies that the principal eigenvalue for $(i\nabla + \kappa h\mathbf{a}_N)^2$ on $\boldsymbol{\mathcal{D}}$ is bounded by $C_1\kappa^2$. We will show that there exists a constant ϕ such that if $h > \max(\frac{1}{\kappa}, \phi\kappa)$, then the principal eigenvalue is greater than $C_1\kappa^2$. It then follows for such κ and h that there are only normal solutions to (1.2).

The corresponding eigenfunctions are expected to take the form of a boundary layer. The following lemma gives a way of measuring to what extent functions can concentrate near the boundary.

For a set $\boldsymbol{\mathcal{O}}$ we define the *τ-neighborhood in $\boldsymbol{\mathcal{O}}$ of $\partial\boldsymbol{\mathcal{O}}$* by

$$\boldsymbol{\mathcal{O}}_\tau = \{\mathbf{x} \in \boldsymbol{\mathcal{O}} : \operatorname{dist}(\mathbf{x}, \partial\boldsymbol{\mathcal{O}}) < \tau\}.$$

LEMMA 2.6. *Let $\boldsymbol{\mathcal{O}}$ be a bounded domain in $\mathbb{R}^n$ with a C^1 boundary. Given $\lambda_0 > 0$ there is a constant $d(\lambda_0, \boldsymbol{\mathcal{O}}) > 0$ such that whenever*

$$\int_{\boldsymbol{\mathcal{O}}} |\nabla f|^2 d\mathbf{x} \leq \lambda^2 \int_{\boldsymbol{\mathcal{O}}} |f|^2 d\mathbf{x} \tag{2.10}$$

for some $f \in H^1(\boldsymbol{\mathcal{O}})$ with $\lambda \geq \lambda_0$, then

$$\frac{1}{2}\int_{\boldsymbol{\mathcal{O}}} |f|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{O}}\setminus\boldsymbol{\mathcal{O}}_{\frac{d}{\lambda}}} |f|^2 d\mathbf{x}. \tag{2.11}$$

Proof. Let $\cup_{k=0}^{N} F_k$ be an open cover for $\overline{\boldsymbol{\mathcal{O}}}$ such that $\overline{F}_0 \subset \boldsymbol{\mathcal{O}}$ and such that for each k, $1 \leq k \leq N$, we have

$$F_k \cap \boldsymbol{\mathcal{O}} = \{(x', x_n) : g_k(x') < x_n < g_k(x') + \delta_1, |x'| < \delta_2\},$$

where δ_1 and δ_2 are positive constants, (x', x_n) are suitably rotated and translated coordinates, and $g_k(x') = x_n$ characterizes $\partial\boldsymbol{\mathcal{O}} \cap F_k$. We can further assume without loss of generality that $g_k(\cdot)$ is defined for $|x'| \leq 2\delta_2$, $|\nabla g_k| < 1$, and

$$\begin{aligned} &\{(x', x_n) : g_k(x') < x_n < g_k(x') + 4\delta_1, |x'| < 2\delta_2\} \subset \boldsymbol{\mathcal{O}}, \\ &\{(x', x_n) : g_k(x') - 4\delta_1 < x_n < g_k(x'), |x'| < 2\delta_2\} \subset \mathbb{R}^n \setminus \overline{\boldsymbol{\mathcal{O}}}. \end{aligned} \tag{2.12}$$

Let $f \in H^1(\boldsymbol{\mathcal{O}})$, $0 \leq t$, $v \leq \delta_1$, and fix $k \geq 1$. We have

$$\begin{aligned} &\int_{\{x_n - g_k(x') = v\} \cap F_k} |f|^2 d\mathbf{s} - \int_{\{x_n - g_k(x') = t\} \cap F_k} |f|^2 d\mathbf{s} \\ &\quad \leq \int_{\{|x'| \leq \delta_2\}} |f^2(x', g_k(x') + v) - f^2(x', g_k(x') + t)| \sqrt{1 + |\nabla g_k|^2}\, dx' \\ &\quad \leq \int_{\boldsymbol{\mathcal{O}} \cap F_k} \left| \frac{\partial (f^2)}{\partial x_n} \right| d\mathbf{x}. \end{aligned}$$

Integrating in t from 0 to δ_1 and then dividing by δ_1 gives

$$\int_{\{x_n - g_k(x') = v\} \cap F_k} |f|^2 d\mathbf{s} \leq \frac{1}{\delta_1} \left(\int_{\boldsymbol{\mathcal{O}}} |f|^2 d\mathbf{x} + 2\delta_1 \int_{\boldsymbol{\mathcal{O}}} |f|\, |\nabla f| d\mathbf{x} \right).$$

Next integrate v from 0 to $\frac{2d}{\lambda}$, where $0 < d < \lambda_0 \delta_1 / 2$ is to be determined. We derive

$$\int_{\{0 < x_n - g_k(x') \leq \frac{2d}{\lambda}\} \cap F_k} |f|^2 d\mathbf{x} \leq \frac{2d}{\lambda \delta_1} \left(\int_{\boldsymbol{\mathcal{O}}} |f|^2 d\mathbf{x} + 2\delta_1 \int_{\mathcal{O}} |f|\, |\nabla f| d\mathbf{x} \right). \tag{2.13}$$

From the assumptions on g_k, for $1 \leq k \leq N$ we have

$$F_k \cap \boldsymbol{\mathcal{O}}_{\frac{d}{\lambda}} \subset F_k \cap \left\{ 0 < x_n - g_k(x') \leq \frac{2d}{\lambda} \right\} \tag{2.14}$$

if $\frac{d}{\lambda}$ is small enough. Indeed, if this is false for some k we can find $\mathbf{x} = (x', x_n)$ such that $|x'| \leq \delta_2$, $x_n > g_k(x') + \frac{2d}{\lambda}$, and $\mathbf{y} = (y', y_n) \in \partial\boldsymbol{\mathcal{O}}$ such that $|\mathbf{x} - \mathbf{y}| < \frac{d}{\lambda}$. If $\frac{d}{\lambda} < \min(\delta_1, \delta_2)$, we have $|x' - y'| < \frac{d}{\lambda} < \delta_2$ implying $|y'| < 2\delta_2$ and $|y_n - x_n| < \frac{d}{\lambda} < \delta_1$. We claim that $y_n = g_k(y')$. In fact,

$$|y_n - g_k(y')| \leq |y_n - x_n| + |x_n - g_k(x')| + |g_k(x') - g_k(y')|,$$

and each term on the right is bounded by δ_1. We have shown this for the first one. This is true for the second since $\mathbf{x} \in F_k \cap \boldsymbol{\mathcal{O}}$. For the last term we use $|\nabla g_k| < 1$ and

$|x' - y'| < \frac{d}{\lambda} < \delta_1$. Thus, $|y_n - g_k(y')| < 3\delta_1$. From (2.12) we see the only possibility for such a $\mathbf{y} \in \partial\mathcal{O}$ is $y_n = g_k(y')$. As a result,

$$|g_k(x') - g_k(y')| \geq |g_k(x') - x_n| - |g_k(y') - x_n| > \frac{2d}{\lambda} - \frac{d}{\lambda} = \frac{d}{\lambda}.$$

On the other hand, since $|\nabla g_k| < 1$ we have $|g_k(x') - g_k(y')| < |x' - y'| < \frac{d}{\lambda}$ and this is a contradiction.

Using (2.13) and (2.14) and summing on k for $1 \leq k \leq N$, we obtain

$$\int_{\mathcal{O}_{\frac{d}{\lambda}}} |f|^2 d\mathbf{x} \leq M_1 \frac{d}{\lambda} \left(\int_{\mathcal{O}} |f|^2 d\mathbf{x} + \lambda \int_{\mathcal{O}} |f|^2 d\mathbf{x} + \lambda^{-1} \int_{\mathcal{O}} |\nabla f|^2 d\mathbf{x} \right),$$

where $M_1 = M_1(\delta_1, N)$.

Using (2.10) we have

$$\int_{\mathcal{O}_{\frac{d}{\lambda}}} |f|^2 d\mathbf{x} \leq M_2 d \int_{\mathcal{O}} |f|^2 d\mathbf{x}$$

where $M_2 = M_2(\delta_1, N, \lambda_0)$. Setting $d = \min(\frac{1}{2M_2}, \frac{\lambda_0\delta_1}{2}, \frac{\lambda_0\delta_2}{2})$, we conclude

$$\int_{\mathcal{O}_{\frac{d}{\lambda}}} |f|^2 d\mathbf{x} \leq \frac{1}{2} \int_{\mathcal{O}} |f|^2 d\mathbf{x}.$$

The assertion (2.11) follows from this inequality. □

We will use the following result from [1] for $B_r(\mathbf{0}) \subset \mathbb{R}^2$.

PROPOSITION 2.7. *There is a continuous function $\sigma(\cdot) : t \in [0, \infty) \to \mathbb{R}$ with $\sigma(t) > 0$ for $t > 0$ for which $\lim_{t\to\infty} \sigma(t)$ exists with $0 < \lim_{t\to\infty} \sigma(t) < 1$, and such that*

$$\int_{B_r(\mathbf{0})} \left| \left(i\nabla + \frac{\omega^2}{2}(-y, x) \right) \zeta \right|^2 d\mathbf{x} \geq \omega^2 \sigma(\omega r) \int_{B_r(\mathbf{0})} |\zeta|^2 d\mathbf{x} \tag{2.15}$$

for all $\zeta \in \mathcal{H}^1(B_r(\mathbf{0}))$ and $\omega \geq 0$.

Indeed in [1, section 2], it is shown that

$$\inf_{\substack{\|\zeta\|_{L^2}=1 \\ \zeta \in W^{1,2}(B_r;\mathbb{C})}} \int_{B_r} \left| \left(i\nabla + \frac{\omega^2}{2}(-y, x) \right) \zeta \right|^2 d\mathbf{x} \equiv \omega^2 \sigma$$

where $\sigma = \sigma(\omega r)$. Furthermore, σ is characterized by $\sigma(t) = \inf_{n\in\mathbf{z}} \sigma(t, n)$ where for each n, $\sigma(t, n)$ is analytic and positive on $0 < t < \infty$. Moreover, $\lim_{t\to 0} \sigma(t, 0) = 0$. In [1, section 6], it is also shown that

$$\sigma(t) = \min_{0\leq n\leq n_0-1} \sigma(t, n) \qquad \text{for } 0 \leq t \leq n_0.$$

As a result, it follows that $\sigma(t)$ is positive and continuous. The $\lim_{t\to\infty} \sigma(t)$ is analyzed in [1, section 6], as well.

Remark. If $\mathbf{b}$ is another vector field such that $\mathbf{b} \in H^1(B_r(\mathbf{0}); \mathbb{R}^2)$ with $\operatorname{curl} \mathbf{b} = \mathbf{e}_3$, then (2.15) is also valid with $\frac{1}{2}(-y, x)$ replaced by $\mathbf{b}$. Indeed, we can define a function

$q \in H^2(B_r(\mathbf{0}))$ such that $\nabla q = \mathbf{b} - \frac{1}{2}(-y,x)$. With this we can define a local gauge transformation

$$\zeta' = \zeta e^{i\omega^2 q}, \qquad \mathbf{b} = \frac{1}{2}(-y,x) + \nabla q,$$

for which $\zeta' \in \mathcal{H}^1(B_r(\mathbf{0}))$ provided $\zeta \in \mathcal{H}^1(B_r(\mathbf{0}))$. Moreover,

$$|(i\nabla + \omega^2 \mathbf{b})\zeta'| = \left| \left(i\nabla + \frac{\omega^2}{2}(-y,x) \right) \zeta \right| \text{ and } |\zeta'| = |\zeta|$$

so that

$$\int_{B_r(\mathbf{0})} |(i\nabla + \omega^2 \mathbf{b})\zeta'|^2 d\mathbf{x} \geq \omega^2 \sigma(\omega r) \int_{B_r(\mathbf{0})} |\zeta'|^2 d\mathbf{x} \tag{2.16}$$

for all $\zeta' \in \mathcal{H}^1(B_r(\mathbf{0}))$, $\mathbf{b} \in H^1(B_r(\mathbf{0}); \mathbb{R}^2)$ such that $\operatorname{curl} \mathbf{b} = \mathbf{e}_3$.

We now derive an estimate similar to (2.16) for $\boldsymbol{\mathcal{D}}$ provided ω is bounded away from zero.

LEMMA 2.8. *Given $m > 0$ there is a constant $C_2 = C_2(m, \boldsymbol{\mathcal{D}})$, $0 < C_2 \leq 1$, such that if $\omega^2 \geq m$, then*

$$C_2 \omega^2 \int_{\boldsymbol{\mathcal{D}}} |\zeta|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \omega^2 \mathbf{b})\zeta|^2 d\mathbf{x} \tag{2.17}$$

for all $\zeta \in \mathcal{H}^1(\boldsymbol{\mathcal{D}})$ and $\mathbf{b} \in H^1(\boldsymbol{\mathcal{D}}; \mathbb{R}^2)$ for which $\operatorname{curl} \mathbf{b} = \mathbf{e}_3$.

Proof. Let $\zeta \in \mathcal{H}^1(\boldsymbol{\mathcal{D}})$ such that $\int_{\boldsymbol{\mathcal{D}}} |\zeta|^2 d\mathbf{x} > 0$ and

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \omega^2 \mathbf{b})\zeta|^2 d\mathbf{x} \leq \omega^2 \int_{\boldsymbol{\mathcal{D}}} |\zeta|^2 d\mathbf{x}$$

for some ω, $\omega^2 \geq m$. If no such ζ exists, then (2.17) is valid with $C_2 = 1$ and we are done. From (2.4) we see $|\nabla |\zeta|\,| \leq |(i\nabla + \omega^2 \mathbf{b})\zeta|$. Thus,

$$\int_{\boldsymbol{\mathcal{D}}} |\nabla |\zeta|\,|^2 d\mathbf{x} \leq \omega^2 \int_{\boldsymbol{\mathcal{D}}} |\zeta|^2 d\mathbf{x}.$$

As a result, we can apply Lemma 2.6 to conclude

$$\frac{1}{2} \int_{\boldsymbol{\mathcal{D}}} |\zeta|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_{\frac{d}{\omega}}} |\zeta|^2 d\mathbf{x}. \tag{2.18}$$

Next we choose a cover for $\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_{\frac{d}{\omega}}$ consisting of a finite collection of disks $\{B_{\frac{d}{\omega}}(\mathbf{x}_k), k = 1, \ldots, N(\omega)\}$, each contained in $\boldsymbol{\mathcal{D}}$ in such a way that $\sum_{k=1}^{N(\omega)} \chi_{B_{\frac{d}{\omega}}(\mathbf{x}_k)} \leq K_1$ where K_1 is independent of ω.

We see

$$\begin{aligned} \int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \omega^2 \mathbf{b})\zeta|^2 d\mathbf{x} &\geq \int_{\cup_{k=1}^N B_{\frac{d}{\omega}}(\mathbf{x}_k)} |(i\nabla + \omega^2 \mathbf{b})\zeta|^2 d\mathbf{x} \\ &\geq K_1^{-1} \sum_{k=1}^N \int_{B_{\frac{d}{\omega}}(\mathbf{x}_k)} |(i\nabla + \omega^2 \mathbf{b})\zeta|^2 d\mathbf{x}. \end{aligned}$$

Using Proposition 2.7, the last term bounds

$$K_1^{-1}\omega^2\sigma(d)\sum_{k=1}^{N}\int_{B_{\frac{d}{\omega}}(\mathbf{x}_k)}|\zeta|^2d\mathbf{x}\geq K_2(d)\omega^2\int_{\boldsymbol{\mathcal{D}}\setminus\boldsymbol{\mathcal{D}}_{\frac{d}{\omega}}}|\zeta|^2d\mathbf{x}\geq\frac{K_2}{2}\omega^2\int_{\boldsymbol{\mathcal{D}}}|\zeta|^2d\mathbf{x}$$

where the final inequality follows from (2.18). Set $C_2 = K_2/2$. This chain of inequalities establishes the lemma. □

We now establish the principal result in this section. Here we prove the existence of an upper critical field $\overline{h}$ and obtain a bound for it as $\kappa\to\infty$ and $\kappa\to 0$.

THEOREM 2.9. *There is a constant $\phi=\phi(\mu_e,\boldsymbol{\mathcal{D}})$ so that if $h>\max(\frac{1}{\kappa},\phi\kappa)$, then any weak solution for* (1.2) *with $n=2$ is normal.*

Proof. Let $(0,\mathbf{a}_N)$ be a normal state for (1.3) and $(\psi,\mathbf{A})$ be a weak solution for (1.2). A state is normal iff its entire gauge equivalence class is normal. Therefore, we can assume without loss of generality that $(\psi,\mathbf{A})$ and $(0,h\mathbf{a}_N)$ satisfy (2.2). Set $\omega^2=h\kappa$. Then $\omega^2\geq 1$ by hypothesis. We apply (2.17) with $m=1$ to derive

$$C_2h\kappa\int_{\boldsymbol{\mathcal{D}}}|\psi|^2d\mathbf{x}\leq\int_{\boldsymbol{\mathcal{D}}}|(i\nabla+h\kappa\mathbf{a}_N)\psi|^2d\mathbf{x},\tag{2.19}$$

and by Lemma 2.5, the right-hand side of (2.19) is bounded by $C_1\kappa^2\int_{\boldsymbol{\mathcal{D}}}|\psi|^2d\mathbf{x}$. Let $\phi=C_1/C_2$. We have

$$h\int_{\boldsymbol{\mathcal{D}}}|\psi|^2d\mathbf{x}\leq\phi\kappa\int_{\boldsymbol{\mathcal{D}}}|\psi|^2d\mathbf{x}.$$

By assumption $h>\phi\kappa$. Hence it must hold that $\int_{\boldsymbol{\mathcal{D}}}|\psi|^2d\mathbf{x}=0$. □

3. Three-dimensional bodies. In this section, we consider a superconducting body given by a bounded domain $\boldsymbol{\mathcal{D}}\subset\mathbb{R}^3$ subjected to a uniform applied field $\mathbf{H}_a = h\mathbf{e}$. We will assume without loss of generality that $\mathbf{e}=\mathbf{e}_3$ throughout this section.

Denote by $\check{H}^1(\mathbb{R}^3)$ the completion of $C_0^\infty(\mathbb{R}^3;\mathbb{R}^3)$ with respect to the norm

$$\|\mathbf{B}\|_{\check{H}^1(\mathbb{R}^3)}=\left(\int_{\mathbb{R}^3}|\nabla\mathbf{B}|^2d\mathbf{x}\right)^{\frac{1}{2}}.$$

One can show elements $\mathbf{B}\in\check{H}^1(\mathbb{R}^3)$ satisfy the following relationships:

$$\|\mathbf{B}\|_{L^6(\mathbb{R}^3;\mathbb{R}^3)}\leq\theta\|\mathbf{B}\|_{\check{H}^1(\mathbb{R}^3)}\tag{3.1}$$

where θ is independent of $\mathbf{B}$ and

$$\|\mathbf{B}\|^2_{\check{H}^1(\mathbb{R}^3)}=\int_{\mathbb{R}^3}(|\operatorname{div}\mathbf{B}|^2+|\operatorname{curl}\mathbf{B}|^2)d\mathbf{x}\tag{3.2}$$

(see [10]).

In order to represent magnetic fields we need the following lemma.

LEMMA 3.1. *Let $\mathbf{g}\in L^2(\mathbb{R}^3;\mathbb{R}^3)$ such that $\operatorname{div}\mathbf{g}=0$ in $\mathcal{D}'(\mathbb{R}^3)$. Then there is a unique $\mathbf{u}\in\check{H}^1(\mathbb{R}^3)$ such that $\operatorname{curl}\mathbf{u}=\mathbf{g}$ and $\operatorname{div}\mathbf{u}=0$.*

Proof. Consider $D_k\Gamma*\mathbf{g}$, where $\Gamma(\mathbf{x})=\Gamma_3(\mathbf{x})=\frac{-1}{3\omega_3|\mathbf{x}|}$ is the Newtonian potential for $\mathbb{R}^3$ and $1\leq k\leq 3$. We claim that $D_k\Gamma*\mathbf{g}\in\check{H}^1(\mathbb{R}^3)$. To see this, let us first assume that $\mathbf{g}$ has bounded support. Then, $D_k\Gamma*g$ exists as a weakly singular integral and

$$|D_k\Gamma*\mathbf{g}|=O(|\mathbf{x}|^{-2})$$

and

$$|\nabla(D_k\Gamma * \mathbf{g})| = O(|\mathbf{x}|^{-3}) \text{ as } |\mathbf{x}| \to \infty.$$

Let $\varphi_R(\mathbf{x})$ be a standard C^∞ cutoff function such that $\varphi_R = 1$ for $|\mathbf{x}| \le R$ and $\varphi_R = 0$ for $|\mathbf{x}| \ge R+1$. It follows directly that $\{\varphi_{R(n)}(D_k\Gamma * \mathbf{g})\}$ is a Cauchy sequence in $\check{H}^1(\mathbb{R}^3)$ that converges to $D_k\Gamma * \mathbf{g}$ pointwise for any sequence $R(n) \to \infty$. Thus, $D_k\Gamma * \mathbf{g} \in \check{H}^1(\mathbb{R}^3)$ assuming $\mathbf{g}$ has bounded support. Finally, by standard L^2-singular integral theory,

$$\|D_k\Gamma * \mathbf{g}\|_{\check{H}^1(\mathbb{R}^3)} \le \|\mathbf{g}\|_{L^2(\mathbb{R}^3;\mathbb{R}^3)}, \tag{3.3}$$

and as a consequence, $D_k\Gamma * \mathbf{g} \in \check{H}^1(\mathbb{R}^3)$ for all $\mathbf{g} \in L^2(\mathbb{R}^3;\mathbb{R}^3)$.

Define $\mathbf{u} : \mathbf{g} \in L^2(\mathbb{R}^3;\mathbb{R}^3) \to \check{H}^1(\mathbb{R}^3)$ by

$$\mathbf{u}(\mathbf{g}) = -(D_2\Gamma * g_3 - D_3\Gamma * g_2, D_3\Gamma * g_1 - D_1\Gamma * g_3, D_1\Gamma * g_2 - D_2\Gamma * g_1).$$

Let $\mathbf{g}_\varepsilon$ be a mollification of $\mathbf{g}$; then $\mathbf{g}_\varepsilon \to \mathbf{g}$ in L^2 as $\varepsilon \to 0$, and $\operatorname{div}\mathbf{g}_\varepsilon = 0$ for each $\varepsilon > 0$. We define $\mathbf{g}_{\varepsilon,R} = \varphi_R\mathbf{g}_\varepsilon$. Note $\operatorname{div}(\mathbf{g}_{\varepsilon,R}) = \nabla\varphi_R \cdot \mathbf{g}_\varepsilon$. If we choose sequences $\varepsilon(n) \to 0$ and $R(n) \to \infty$ as $n \to \infty$, then $\mathbf{g}_n = \varphi_{R(n)}\mathbf{g}_{\varepsilon(n)} \in C_0^\infty(\mathbb{R}^3;\mathbb{R}^3)$, $\mathbf{g}_n \to \mathbf{g}$ in L^2, and $\operatorname{div}\mathbf{g}_n \to 0$ in L^2 as $n \to \infty$.

Set $\mathbf{w}_n = \Gamma * \mathbf{g}_n$. These are well defined since $\mathbf{g}_n$ have bounded support. Using (3.3), we see

$$\operatorname{curl}\mathbf{w}_n = -\mathbf{u}(\mathbf{g}_n) \to -\mathbf{u}(\mathbf{g}) \text{ in } \check{H}^1 \text{ as } n \to \infty.$$

Consider

$$\operatorname{curl}\mathbf{u}(\mathbf{g}_n) = -\operatorname{curl}\operatorname{curl}\mathbf{w}_n = \Delta\mathbf{w}_n - \nabla(\operatorname{div}\mathbf{w}_n) = \mathbf{g}_n - \nabla(\operatorname{div}\mathbf{w}_n). \tag{3.4}$$

We know $\nabla(\operatorname{div}\mathbf{w}_n) = \nabla\Gamma * (\operatorname{div}\mathbf{g}_n) \to \mathbf{0}$ in $\check{H}^1(\mathbb{R}^3)$ as $n \to \infty$ since $\operatorname{div}\mathbf{g}_n \to 0$ in L^2. Thus, using (3.1) we conclude $\nabla(\operatorname{div}\mathbf{w}_n) \to \mathbf{0}$ in L^6, as $n \to \infty$. Furthermore, we have $\operatorname{curl}\mathbf{u}(\mathbf{g}_n) \to \operatorname{curl}\mathbf{u}(\mathbf{g})$ and $\mathbf{g}_n \to \mathbf{g}$ in L^2 as $n \to \infty$. As a consequence $\operatorname{curl}\mathbf{u}(\mathbf{g}) = \mathbf{g}$ in $\mathbb{R}^3$.

Since $\mathbf{u}(\mathbf{g}_n) = -\operatorname{curl}\mathbf{w}_n$ we have $\operatorname{div}\mathbf{u}(\mathbf{g}_n) = 0$, which implies $\operatorname{div}\mathbf{u}(\mathbf{g}) = 0$. Finally, using (3.2) we see that $\mathbf{u}$ is unique in $\check{H}^1(\mathbb{R}^3)$. ∎

We can apply the preceding lemma to characterize weak solutions.

LEMMA 3.2. *Let $(\zeta, \mathbf{B})$ be a weak solution to (1.2). Then there is a gauge equivalent solution $(\psi, \mathbf{A})$ such that $\operatorname{div}\mathbf{A} = 0$ and $(\mathbf{A} - \frac{\mu_e h}{2}(-y, x, 0)) \in \check{H}^1(\mathbb{R}^3)$. Moreover, if $(\tilde{\psi}, \tilde{\mathbf{A}})$ is another such solution, then $\tilde{\psi} = a\psi$ for some $a \in \mathbb{C}, |a| = 1$, and $\tilde{\mathbf{A}} = \mathbf{A}$.*

Proof. Set $\mathbf{g} = (\operatorname{curl}\mathbf{B} - \mu_e h\mathbf{e}_3) \in L^2(\mathbb{R}^3;\mathbb{R}^3)$. From the previous lemma there is a unique element $\mathbf{u} \in \check{H}^1(\mathbb{R}^3)$ such that $\operatorname{curl}\mathbf{u} = \mathbf{g}$ and $\operatorname{div}\mathbf{u} = 0$. Therefore, we find $\mathbf{A} = \mathbf{u} + \frac{\mu_e h}{2}(-y, x, 0)$. ∎

We now characterize the normal state in three dimensions.

LEMMA 3.3. *There is a unique normal state satisfying (1.3) such that $(\mathbf{a}_N - \frac{\mu_e}{2}(-y, x, 0)) \in \check{H}^1(\mathbb{R}^3)$ and $\operatorname{div}\mathbf{a}_N = 0$.*

Proof. Consider the strictly convex functional

$$E(\mathbf{b}) = G(0, \mathbf{b}) + \int_{\mathbb{R}^3} (\operatorname{div}\mathbf{b})^2 d\mathbf{x} = \int_{\boldsymbol{\mathcal{D}}} \frac{1}{2}\, d\mathbf{x} + \int_{\mathbb{R}^3} \left(\mu \left| \frac{1}{\mu} \operatorname{curl}\mathbf{b} - e_3 \right|^2 + (\operatorname{div}\mathbf{b})^2 \right) d\mathbf{x}$$

for the class $\boldsymbol{\mathcal{S}} = \{\mathbf{b} : (\mathbf{b} - \frac{\mu_e}{2}(-y, x, 0)) \in \check{H}^1(\mathbb{R}^3)\}$. A unique equilibrium exists which also minimizes $E(\cdot)$. Let $\widetilde{\mathbf{b}}$ be this equilibrium. If $\operatorname{div}\widetilde{\mathbf{b}} \neq 0$ then by Lemma 3.2 we can find another vector field $\widetilde{\widetilde{\mathbf{b}}} \in \boldsymbol{\mathcal{S}}$ such that $\operatorname{curl}\widetilde{\widetilde{\mathbf{b}}} = \operatorname{curl}\widetilde{\mathbf{b}}$ and $\operatorname{div}\widetilde{\widetilde{\mathbf{b}}} = 0$. This would imply $E(\widetilde{\widetilde{\mathbf{b}}}) < E(\widetilde{\mathbf{b}})$ which is impossible. Thus, $\operatorname{div}\widetilde{\mathbf{b}} = 0$. It follows that $\widetilde{\mathbf{b}}$ satisfies (1.3) and as a result a normal state $(0, \mathbf{a}_N)$ exists. Conversely, a normal state satisfying the hypothesis is an equilibrium for $E(\cdot)$, and so $\mathbf{a}_N$ is unique. ∎

Recall that the induction $\operatorname{curl}\mathbf{a}_N$, not $\mathbf{a}_N$, is the physically relevant quantity. Below we show it is uniquely determined.

LEMMA 3.4. *Let $(0, \mathbf{a}_N)$ be a normal state. Then $\operatorname{curl}\mathbf{a}_N$ is uniquely determined, $\operatorname{curl}\mathbf{a}_N$ is harmonic in $\mathbb{R}^3\backslash\partial\boldsymbol{\mathcal{D}}$ and*

$$\operatorname{curl}\mathbf{a}_N \in C^{1,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{1,\alpha}(\boldsymbol{\mathcal{D}}^c).$$

Moreover, if $\operatorname{div}\mathbf{a}_N = 0$, then

$$\mathbf{a}_N \in C^{1,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{1,\alpha}(\boldsymbol{\mathcal{D}}^c).$$

Proof. Using Lemma 3.2 we see any normal state is gauge equivalent to the normal state described in Lemma 3.3. Since a gauge transformation leaves the curl of a vector field invariant, we conclude that $\operatorname{curl}\mathbf{a}_N$ is uniquely determined for solutions to (1.3).

We can use the first equation in (1.3) to prove that there exists a function $p \in H^1_{\text{loc}}(\mathbb{R}^3)$ such that $\operatorname{curl}\mathbf{a}_N = \mu\nabla p$. Since

$$\operatorname{div}(\mu\nabla p) = 0 \text{ in } \mathbb{R}^3, \tag{3.5}$$

and μ is constant on the components of $\mathbb{R}^3 \setminus \partial\boldsymbol{\mathcal{D}}$, the function p (and thus $\operatorname{curl}\mathbf{a}_N$) is harmonic in each component. We apply the results from [11, Chapter 5, Section 4] to the solution p for (3.5), to derive that $p \in C^{2,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{2,\alpha}(\boldsymbol{\mathcal{D}}^c)$ and as a consequence, $\operatorname{curl}\mathbf{a}_N \in C^{1,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{1,\alpha}(\boldsymbol{\mathcal{D}}^c)$.

Assume $\operatorname{div}\mathbf{a}_N = 0$. Let $\boldsymbol{\mathcal{U}}$ be an open neighborhood of $\overline{\boldsymbol{\mathcal{D}}}$ and consider $\mathbf{w} \in H^2(\boldsymbol{\mathcal{U}};\mathbb{R}^3)$ such that $\Delta\mathbf{w} = \operatorname{curl}\mathbf{a}_N$ in $\boldsymbol{\mathcal{U}}$. From [11, Chapter 5], we have $\mathbf{w} \in C^{2,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{2,\alpha}(\boldsymbol{\mathcal{U}} \setminus \boldsymbol{\mathcal{D}})$. The identity $\operatorname{curl}(\operatorname{curl}\mathbf{w}) + \Delta\mathbf{w} = \nabla(\operatorname{div}\mathbf{w})$ in $\boldsymbol{\mathcal{U}}$ yields

$$\operatorname{curl}(\operatorname{curl}\mathbf{w} + \mathbf{a}_N) = \nabla(\operatorname{div}\mathbf{w}),$$

from which we obtain

$$\operatorname{curl}\operatorname{curl}(\operatorname{curl}\mathbf{w} + \mathbf{a}_N) = \mathbf{0} \text{ in } \mathcal{D}'(\boldsymbol{\mathcal{U}}).$$

By hypothesis $\operatorname{div}(\operatorname{curl}\mathbf{w} + \mathbf{a}_N) = 0$. Whence, from the identity above

$$-\Delta(\operatorname{curl}\mathbf{w} + \mathbf{a}_N) = \operatorname{curl}\operatorname{curl}(\operatorname{curl}\mathbf{w} + \mathbf{a}_N) = \mathbf{0} \text{ in } \mathcal{D}'(\boldsymbol{\mathcal{U}}).$$

This implies $(\operatorname{curl}\mathbf{w} + \mathbf{a}_N) \in C^\infty(\boldsymbol{\mathcal{U}})$, and we conclude

$$\mathbf{a}_N \in C^{1,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{1,\alpha}(\boldsymbol{\mathcal{D}}^c). \qquad ∎$$

Consider the case $\mu \equiv 1$. Given $\mathbf{e}$, we can find a linear function $\mathbf{a}(\mathbf{x})$ such that $\operatorname{curl}\mathbf{a} \equiv \mathbf{e}$. Clearly $\mathbf{a}$ satisfies (1.3). It follows from the previous lemma then that $\operatorname{curl}\mathbf{a}_N \equiv \mathbf{e}$ when $\mu \equiv 1$.

We now derive a Sobolev estimate analogous to Lemma 2.3.

LEMMA 3.5. *Let* $(\zeta, \mathbf{B})$ *be a weak solution to* (1.2). *Let* $(\psi, \mathbf{A})$ *be the gauge equivalent solution found in Lemma* 3.2 *and* $(0, h\mathbf{a}_N)$ *be the normal state found in Lemma* 3.3. *Then there is a constant* C_0 *depending only on* $\boldsymbol{\mathcal{D}}$ *such that*

$$\int_{\boldsymbol{\mathcal{D}}} |\mathbf{A} - h\mathbf{a}_N|^2 d\mathbf{x} \le C_0 \int_{\mathbb{R}^3} |\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)|^2 d\mathbf{x}.$$

Proof. Using (3.1) and (3.2) we see

$$\|\mathbf{A} - h\mathbf{a}_N\|_{L^6(\mathbb{R}^3;\mathbb{R}^3)} \le \theta \|\nabla(\mathbf{A} - h\mathbf{a}_N)\|_{L^2(\mathbb{R}^3;\mathbb{R}^3)} = \theta \|\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)\|_{L^2(\mathbb{R}^3;\mathbb{R}^3)}.$$

Since $\boldsymbol{\mathcal{D}}$ is bounded we have

$$\|\mathbf{A} - h\mathbf{a}_N\|_{L^2(\boldsymbol{\mathcal{D}};\mathbb{R}^3)} \le M(\boldsymbol{\mathcal{D}}) \|\mathbf{A} - h\mathbf{a}_N\|_{L^6(\boldsymbol{\mathcal{D}};\mathbb{R}^3)}$$

and the lemma follows. ∎

We proceed in deriving the three-dimensional counterpart to Lemma 2.5.

LEMMA 3.6. *Let* $(\psi, \mathbf{A})$ *and* $(0, h\mathbf{a}_N)$ *be as in Lemma* 3.5. *Then there is a constant* $C_1 = C_1(\boldsymbol{\mathcal{D}}, \mu_e)$ *so that*

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + h\kappa \mathbf{a}_N)\psi|^2 d\mathbf{x} \le C_1 \kappa^2 \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x}.$$

Proof. We proceed just as in Lemma 2.5 to obtain

$$\int_{\mathbb{R}^3} \frac{1}{\mu} [\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N) \cdot \operatorname{curl} \mathbf{B}] d\mathbf{x} \le \varepsilon^{-1} \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} + \varepsilon \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 |\mathbf{B}|^2 d\mathbf{x}$$

for any $\varepsilon > 0$ and $\mathbf{B} \in H^1(\mathbb{R}^3;\mathbb{R}^3)$ with bounded support. However, since $\mathbf{A} - h\mathbf{a}_N \in \check{H}^1(\mathbb{R}^3)$, we can take $\mathbf{B} = \mathbf{B}_j \to \mathbf{A} - h\mathbf{a}_N$ in $\check{H}^1(\mathbb{R}^3)$ as $j \to \infty$. As a result, we have

$$\int_{\mathbb{R}^3} \frac{1}{\mu} |\operatorname{curl}(\mathbf{A} - h\mathbf{a}_N)|^2 d\mathbf{x} \le \varepsilon^{-1} \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} + \varepsilon \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 |\mathbf{A} - h\mathbf{a}_N|^2 d\mathbf{x}.$$

The remainder of the proof is just as before. ∎

We next give a three-dimensional analogue for the eigenvalue estimate from [1].

Let $\mathbf{v} \in \mathbb{R}^3 \setminus \{\mathbf{0}\}$ such that $|\mathbf{v}| = 1$ and $\mathbf{x}_0 \in \mathbb{R}^3$. Let $T(\mathbf{x}_0, r, \mathbf{v})$ be a cylinder with central axis parallel to $\mathbf{v}$, height $2r$, and middle cross section the disk of radius r with center $\mathbf{x}_0$.

LEMMA 3.7. *Let* $\mathbf{b} \in H^1(T(\mathbf{x}_0, r, \mathbf{v}); \mathbb{R}^3)$ *such that* $\operatorname{curl} \mathbf{b} = \mathbf{v}$. *Then*

$$\int_T |(i\nabla + \omega^2 \mathbf{b})\zeta|^2 d\mathbf{x} \ge \omega^2 \sigma(\omega r) \int_T |\zeta|^2 d\mathbf{x} \tag{3.6}$$

for all $\zeta \in \mathcal{H}^1(T)$ *where* $\sigma(\cdot)$ *is as in Proposition* 2.7.

Proof. We first transfer the problem to

$$T(\mathbf{0}, r, \mathbf{e}_3) = B_r(\mathbf{0}) \times (-r, r).$$

Let $Q \in SO(3)$ such that $\mathbf{e}_3 = Q\mathbf{v}$, and set $\mathbf{y}(\mathbf{x}) = Q(\mathbf{x} - \mathbf{x}_0)$. Then

$$\mathbf{y} : \mathbf{x} \in T(\mathbf{x}_0, r, \mathbf{v}) \to T(\mathbf{0}, r, \mathbf{e}_3).$$

Given $\zeta \in \mathcal{H}^1(T(\mathbf{x}_0, r, \mathbf{v}))$ we define $\xi(\mathbf{y}) = \zeta(\mathbf{x}(\mathbf{y}))$. Then

$$(i\nabla_{\mathbf{y}} + \omega^2 \mathbf{b} Q^t)\xi(\mathbf{y}) = (i\nabla_{\mathbf{x}} + \omega^2 \mathbf{b})\zeta(\mathbf{x}) Q^t.$$

By changing variables we see that (3.6) is equivalent to showing the following inequality:

$$\int_{T(\mathbf{0},r,\mathbf{e}_3)} |(i\nabla + \omega^2 \mathbf{b} Q^t)\xi|^2 d\mathbf{y} \geq \omega^2 \sigma(\omega r) \int_{T(\mathbf{0},r,\mathbf{e}_3)} |\xi|^2 d\mathbf{y}.$$

For any $\mathbf{w} \in \mathbb{R}^3$, we have

$$Q\mathbf{w} \cdot \operatorname{curl}_{\mathbf{y}}(\mathbf{b}Q^t) = \det \begin{bmatrix} \mathbf{w}^t Q^t \\ \nabla_{\mathbf{y}} \\ \mathbf{b} Q^t \end{bmatrix} = \det \begin{bmatrix} \mathbf{w}^t Q^t \\ \nabla_{\mathbf{x}} Q^t \\ \mathbf{b} Q^t \end{bmatrix} = \det \begin{bmatrix} \mathbf{w}^t \\ \nabla_{\mathbf{x}} \\ \mathbf{b} \end{bmatrix} = \mathbf{w} \cdot \operatorname{curl}_{\mathbf{x}} \mathbf{b}.$$

Therefore, $\operatorname{curl}_{\mathbf{y}}(\mathbf{b}Q^t) = Q(\operatorname{curl}_{\mathbf{x}} \mathbf{b}) = Q\mathbf{v} = \mathbf{e}_3$.

By changing the gauge if necessary, we can assume $\mathbf{b}Q^t = \frac{1}{2}(-y, x, 0)$. Then

$$\begin{aligned}
\int_{T(\mathbf{0},r,\mathbf{e}_3)} |(i\nabla + \omega^2 \mathbf{b} Q^t)\xi|^2 d\mathbf{y} &\geq \int_{T(\mathbf{0},r,\mathbf{e}_3)} |(i(D_x, D_y, 0) + \omega^2 \mathbf{b} Q^t)\xi|^2 d\mathbf{y} \\
&= \int_{-r}^{r} \int_{B_r} \left| \left(i(D_x, D_y, 0) + \frac{\omega^2}{2}(-y, x, 0) \right) \xi(x, y, z) \right|^2 dx\, dy\, dz \\
&\geq \int_{-r}^{r} \omega^2 \sigma(\omega r) \int_{B_r} |\xi(x, y, z)|^2 dx\, dy\, dz \\
&= \omega^2 \sigma(\omega r) \int_{T(\mathbf{0},r,\mathbf{e}_3)} |\xi|^2 d\mathbf{y},
\end{aligned}$$

where we have applied Proposition 2.7 for each $-r \leq z \leq r$. ∎

We go on to prove the three-dimensional counterpart to the eigenvalue estimate in Lemma 2.8.

LEMMA 3.8. *Let* $(0, \mathbf{a}_N)$ *be the normal state from Lemma* 3.3. *Assume* $\operatorname{curl} \mathbf{a}_N \neq \mathbf{0}$ *in* $\overline{\boldsymbol{\mathcal{D}}}$. *Then there exist constants* $m \geq 1$ *and* $0 < C_2 \leq 1$ *so that if* $\omega^2 \geq m$, *it holds that*

$$C_2 \omega^2 \int_{\boldsymbol{\mathcal{D}}} |\zeta|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \omega^2 \mathbf{a}_N)\zeta|^2 d\mathbf{x} \tag{3.7}$$

for all $\zeta \in \mathcal{H}^1(\boldsymbol{\mathcal{D}})$.

Proof. We argue as in Lemma 2.8. There exists a constant $d > 0$ so that given $\xi \in \mathcal{H}^1(\boldsymbol{\mathcal{D}})$ and $\omega \geq 1$ either (3.7) is true with $C_2 = 1$ and $\xi = \zeta$ or

$$\frac{1}{2} \int_{\boldsymbol{\mathcal{D}}} |\xi|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_{\frac{d}{\omega}}} |\xi|^2 d\mathbf{x}. \tag{3.8}$$

Assume the latter. In this case we cover $\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_{\frac{d}{\omega}}$ by a family of cylinders $\{T_k : k = 1, \ldots, N(\omega)\}$ such that $T_k = T(\mathbf{x}_k, \frac{d}{2\omega}, \operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)/|\operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)|)$ with $T_k \subset \boldsymbol{\mathcal{D}}$ for each k and $\sum_{k=1}^{N} \chi_{T_k} \leq K$ where K is independent of ω for $\omega \geq 1$. As a consequence,

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \omega^2 \mathbf{a}_N)\xi|^2 d\mathbf{x} \geq K^{-1} \sum_{k=1}^{N} \int_{T_k} |(i\nabla + \omega^2 \mathbf{a}_N)\xi|^2 d\mathbf{x}. \tag{3.9}$$

In each T_k we write $\mathbf{a}_N(\mathbf{x}) = \ell_k(\mathbf{x}) + q_k(\mathbf{x})$ where $\ell_k(\mathbf{x}) = \mathbf{a}_N(\mathbf{x}_k) + \nabla \mathbf{a}_N(\mathbf{x}_k) \cdot (\mathbf{x} - \mathbf{x}_k)$. Note $\operatorname{curl} \ell_k(\mathbf{x}) = \operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)$. Using (2.9) for each k we obtain

$$\int_{T_k} |(i\nabla + \omega^2 \mathbf{a}_N)\xi|^2 d\mathbf{x} \geq \frac{1}{2} \int_{T_k} |(i\nabla + \omega^2 \ell_k)\xi|^2 d\mathbf{x} - \omega^4 \int_{T_k} |q_k|^2 |\xi|^2 d\mathbf{x}.$$

From Lemma 3.7 we have

$$\begin{aligned}\int_{T_k} |(i\nabla + \omega^2 \ell_k)\xi|^2 d\mathbf{x} &\geq \omega^2 |\operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)| \sigma\left(|\operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)|^{\frac{1}{2}} \frac{d}{2} \right) \int_{T_k} |\xi|^2 d\mathbf{x} \\ &\geq \omega^2 M_0 \int_{T_k} |\xi|^2 d\mathbf{x},\end{aligned}$$

where $M_0 > 0$ depends on $\inf_{\boldsymbol{\mathcal{D}}} |\operatorname{curl} \mathbf{a}_N| > 0$ and the structure of $\sigma(\cdot)$ (see Proposition 2.7).

Since $\mathbf{a}_N \in C^{1,\alpha}(\overline{\boldsymbol{\mathcal{D}}})$ we have

$$|q_k| \leq M_1 (\operatorname{diam} T_k)^{1+\alpha} \leq M_2 \omega^{-1-\alpha} \quad \text{where } M_2 \text{ is independent of } k.$$

As a result, we see for each k that

(3.10)
$$\int_{T_k} |(i\nabla + \omega^2 \mathbf{a}_N)\xi|^2 d\mathbf{x} \geq \left(\frac{M_0}{2} \omega^2 - M_2^2 \omega^{2-2\alpha} \right) \int_{T_k} |\xi|^2 d\mathbf{x} \geq \frac{M_0}{4} \omega^2 \int_{T_k} |\xi|^2 d\mathbf{x},$$

provided $\omega^2 \geq m = m(\boldsymbol{\mathcal{D}}, \mu_e)$ sufficiently large.

From (3.9) and (3.10) then

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + \omega^2 \mathbf{a}_N)\xi|^2 d\mathbf{x} \geq M_3 \omega^2 \int_{\cup_{k=1}^N T_k} |\xi|^2 d\mathbf{x},$$

for some $M_3 > 0$ independent of the cover. Using $\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_{\frac{d}{\omega}} \subset \cup_{k=1}^N T_k$ and (3.8) we derive

$$\int_{\cup_{k=1}^N T_k} |\xi|^2 d\mathbf{x} \geq \int_{\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_{\frac{d}{\omega}}} |\xi|^2 d\mathbf{x} \geq \frac{1}{2} \int_{\boldsymbol{\mathcal{D}}} |\xi|^2 d\mathbf{x}.$$

Setting $C_2 = \frac{M_3}{2}$ we have our lemma. ∎

The following theorem is proved in the same manner as Theorem 2.9. We establish the existence of $\overline{h}$ and derive an upper bound for it provided $\operatorname{curl} \mathbf{a}_N$ does not vanish on $\overline{\boldsymbol{\mathcal{D}}}$.

THEOREM 3.9. *Assume that* $\operatorname{curl} \mathbf{a}_N \neq \mathbf{0}$ *in* $\overline{\boldsymbol{\mathcal{D}}}$. *There are constants* m *and* ϕ, *depending on* $\boldsymbol{\mathcal{D}}$ *and* μ_e, *so that if* $h > \max(\frac{m}{\kappa}, \phi\kappa)$ *then any weak solution to* (1.2) *with* $n = 3$ *is normal.*

For the case $\mu \equiv 1$, we have $|\operatorname{curl} \mathbf{a}_N| \equiv 1$ and we can recover the following result.

COROLLARY 3.10. *If* $\mu_e = 1$ *then there exist constants* m *and* ϕ *depending on* $\boldsymbol{\mathcal{D}}$ *so that if* $h > \max(\frac{m}{\kappa}, \phi\kappa)$; *then any weak solution to* (1.2) *with* $n = 3$ *is normal.*

In general, one does not know if $\operatorname{curl} \mathbf{a}_N$ vanishes somewhere in $\overline{\boldsymbol{\mathcal{D}}}$ or not. Nevertheless, since $\operatorname{curl} \mathbf{a}_N$ is harmonic in $\boldsymbol{\mathcal{D}}$ it can vanish only on a small set.

LEMMA 3.11. *Let* $(0, \mathbf{a}_N)$ *be a normal state. Then* $\mathcal{L}^3(\{\mathbf{x} \in \boldsymbol{\mathcal{D}} : \operatorname{curl} \mathbf{a}_N(\mathbf{x}) = \mathbf{0}\}) = 0$.

Proof. Since $\operatorname{curl} \mathbf{a}_N$ is harmonic in $\boldsymbol{\mathcal{D}}$ either $\mathcal{L}^3(\{\mathbf{x} \in \boldsymbol{\mathcal{D}} : \operatorname{curl} \mathbf{a}_N(\mathbf{x}) = \mathbf{0}\}) = 0$ or $\operatorname{curl} \mathbf{a}_N \equiv \mathbf{0}$ in $\boldsymbol{\mathcal{D}}$. From Lemma 3.4, we know there is a function $p \in C^{2,\alpha}(\overline{\boldsymbol{\mathcal{D}}}) \cap C^{2,\alpha}(\boldsymbol{\mathcal{D}}^c) \cap C(\mathbb{R}^3)$ such that

$$\operatorname{curl} \mathbf{a}_N = \mu \nabla p \text{ in } \mathbb{R}^3.$$

Hence, p satisfies

$$\begin{cases} \operatorname{div}(\mu \nabla p) = 0 & \text{in } \mathbb{R}^3, \\ (\nabla p - \mathbf{e}_3) \in L^2(\mathbb{R}^3). \end{cases} \tag{3.11}$$

Assume that $\operatorname{curl} \mathbf{a}_N \equiv \mathbf{0}$ in $\boldsymbol{\mathcal{D}}$. Then $p = \text{constant} = p_0$ in $\overline{\boldsymbol{\mathcal{D}}}$. It follows from the first equation in (3.11) that p solves

$$\begin{aligned} \Delta p = 0 \qquad & \text{in} \quad \boldsymbol{\mathcal{D}}^c, \\ p = p_0, \frac{\partial p}{\partial n} = 0 \quad & \text{on } \partial \boldsymbol{\mathcal{D}}^c, \end{aligned}$$

where $\mathbf{n}$ is the exterior normal to $\partial \boldsymbol{\mathcal{D}}$. The Cauchy problem has the unique solution $p = p_0$. This would contradict the second equation in (3.11). □

To conclude, we show there is a finite upper critical field for each κ.

THEOREM 3.12. *Let κ, μ_e, and γ be fixed. There is a constant $\overline{h} = \overline{h}(\kappa, \mu_e, \gamma, \boldsymbol{\mathcal{D}})$ so that if $h > \overline{h}$ then any weak solution to* (1.2) *with $n = 3$ is normal.*

Proof. Let $(0, \mathbf{a}_N)$ be as in Lemma 3.3. Assume there exists a sequence $\{(\psi_j, \mathbf{A}_j)\}$ where for each j the pair is as in Lemma 3.5 solving (1.2) with $h = h_j$ for which $\lim_{j\to\infty} h_j = \infty$ and $\int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 dx > 0$.

Set $\varphi_j(x) = |\psi_j(x)|/\|\psi_j\|_{L^2(\boldsymbol{\mathcal{D}})}$. From (2.6) we have

$$\int_{\boldsymbol{\mathcal{D}}} |\nabla \varphi_j|^2 d\mathbf{x} \leq \kappa^2,$$

and we can find a subsequence $\varphi_j \to \varphi_0$ in $L^2(\boldsymbol{\mathcal{D}})$ as $j \to \infty$ with $\|\varphi_0\|_{L^2(\boldsymbol{\mathcal{D}})} = 1$.

From Lemma 3.11 the set $\mathbf{Q} = \{\mathbf{x} \in \overline{\boldsymbol{\mathcal{D}}} : \operatorname{curl} \mathbf{a}_N(\mathbf{x}) = \mathbf{0}\}$ is a closed set of measure zero. It follows that there exists a ball $B_{2r} \subset \boldsymbol{\mathcal{D}} \setminus \mathbf{Q}$ such that $\int_{B_r} |\varphi_0|^2 d\mathbf{x} \equiv 2\delta > 0$ and $\inf_{B_{2r}} |\operatorname{curl} \mathbf{a}_N| > 0$. Note for j sufficiently large we have

$$\int_{B_r} |\psi_j|^2 d\mathbf{x} \geq \delta \int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 d\mathbf{x}.$$

For each j we cover B_r by a finite family of cylinders,

$$\{T(\mathbf{x}_k, \omega_j^{-1}, \operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)/|\operatorname{curl} \mathbf{a}_N(\mathbf{x}_k)|)\} = \{T_{kj}, \quad 1 \leq k \leq N(j)\}$$

such that $\omega_j^2 = h_j \kappa$, $\mathbf{x}_k \in B_r$ and such that the family has overlap of at most K uniformly in $\mathbf{x} \in \boldsymbol{\mathcal{D}}$ independent of j. For j sufficiently large, each $T_{kj} \subset B_{2r}$ and as in Lemma 3.8 we derive,

$$\int_{T_{kj}} |(i\nabla + \omega_j^2 \mathbf{a}_N)\psi_j|^2 d\mathbf{x} \geq \omega_j^2 M_0 \int_{T_{kj}} |\psi_j|^2 d\mathbf{x},$$

where $M_0 > 0$ depends on $\inf_{B_{2r}} |\operatorname{curl} \mathbf{a}_N|$ and $\|\mathbf{a}_N\|_{C^{1,\alpha}(\overline{\boldsymbol{\mathcal{D}}})}$ but not on j. Whence, we can write

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + h_j\kappa \mathbf{a}_N)\psi_j|^2 d\mathbf{x} \geq K^{-1} \sum_{k=1}^{N(j)} \int_{T_{kj}} |(i\nabla + h_j\kappa \mathbf{a}_N)\psi_j|^2 d\mathbf{x}$$
$$\geq M_1 h_j \kappa \int_{\cup_{k=1}^{N(j)} T_{kj}} |\psi_j|^2 d\mathbf{x} \geq M_1 h_j \kappa \int_{B_r} |\psi_j|^2 d\mathbf{x} \geq M_2 h_j \kappa \int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 d\mathbf{x},$$

where M_1 and M_2 are positive and depend on $\inf_{B_{2r}} |\operatorname{curl} \mathbf{a}_N|$, $\mathbf{a}_N$ and δ. From Lemma 3.6 we know

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + h_j\kappa \mathbf{a}_N)\psi_j|^2 d\mathbf{x} \leq C_1 \kappa^2 \int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 d\mathbf{x},$$

which leads to

$$M_2 h_j \kappa \int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 d\mathbf{x} \leq C_1 \kappa^2 \int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 d\mathbf{x}$$

for all j sufficiently large. Since $h_j \to \infty$ as $j \to \infty$, we must have $\int_{\boldsymbol{\mathcal{D}}} |\psi_j|^2 d\mathbf{x} = 0$ for j large and this is a contradiction. ∎

4. Estimates for small κ. In this section we consider $\overline{h}$ for small κ in cases where $\operatorname{curl} \mathbf{a}_N \equiv \mathbf{e}$ in $\boldsymbol{\mathcal{D}}$.

THEOREM 4.1. *Let $n = 2$ with $\mu_e > 0$ or $n = 3$ with $\mu_e = 1$. Then $\overline{h} = O(1)$ as $\kappa \downarrow 0$.*

Proof. We consider the case $n = 3$. The argument for $n = 2$ is identical.

Let $\kappa \leq 1$ and assume (ψ, A) solves (1.2) with $\psi \not\equiv 0$. From (2.6) we have

$$\int_{\boldsymbol{\mathcal{D}}} |\nabla |\psi|\,|^2 d\mathbf{x} \leq \kappa^2 \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x}.$$

If we apply Lemma 2.6 with $\lambda = \lambda_0 = 1$, we find there is a constant $d > 0$ so that

$$\frac{1}{2} \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_d} |\psi|^2 d\mathbf{x}.$$

We take r, $0 < r < d$, to be determined and cover $\boldsymbol{\mathcal{D}} \setminus \boldsymbol{\mathcal{D}}_d$ by a family of cylinders $\{T(\mathbf{x}_k, r, \mathbf{e})\}$ such that the cylinders have finite overlap independent of r and each is contained in $\boldsymbol{\mathcal{D}}$. Then arguing just as in Lemma 2.8 there exists a constant $C_2 > 0$ for which

$$C_2 h\kappa\ \sigma((h\kappa)^{\frac{1}{2}} r) \int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} \leq \int_{\boldsymbol{\mathcal{D}}} |(i\nabla + h\kappa \mathbf{a}_N)\psi|^2 d\mathbf{x}.$$

Applying Lemma 3.6 to the right-hand side, and recalling that $\int_{\boldsymbol{\mathcal{D}}} |\psi|^2 d\mathbf{x} > 0$ the following inequality holds:

$$C_2 h\kappa\ \sigma((h\kappa)^{\frac{1}{2}} r) \leq C_1 \kappa^2. \tag{4.1}$$

From Corollary 3.10 we have $h\kappa \leq M_1^2$ for some $M_1 < \infty$ for all $\kappa \leq 1$. We now choose r so that $M_1 r \leq 2^{\frac{1}{2}}$. The reason for this choice is that if $0 < t \leq 2^{\frac{1}{2}}$, then $\sigma(t)$ is the principal eigenvalue for the Sturm–Liouville problem:

$$g''(s) + \frac{g'(s)}{s} - \frac{s^2 g(s)}{4} = -\sigma(t) g(s) \qquad \text{for } 0 < s < t,$$
$$g'(t) = 0 \text{ and } g \text{ is bounded}$$

(see [1, Section 6]). As such, from [6, Proposition 3.4], we have $\sigma(t) = \frac{t^2}{8} + o(t^2)$ as $t \to 0$. It follows that there is a constant $M_2 > 0$ so that

$$M_2 t^2 \leq \sigma(t) \qquad \text{for } 0 \leq t \leq 2^{\frac{1}{2}}.$$

Combining this estimate with (4.1), we derive

$$h^2 \leq \frac{C_1}{C_2 M_2 r^2} \qquad \text{for } 0 < \kappa \leq 1.$$

Since this holds for all superconducting solutions we conclude

$$\overline{h}^2 \leq \frac{C_1}{C_2 M_2 r^2} \qquad \text{for } 0 < \kappa \leq 1. \qquad \square$$

The length scale for variations in superconducting solutions is $\frac{1}{\kappa}$. Because of this, it is of interest to consider domains with dimensions comparable to $\frac{1}{\kappa}$. To this end, let $\boldsymbol{\mathcal{D}} \subseteq \mathbb{R}^n$ and define the dilated domain

$$\boldsymbol{\mathcal{D}}(\kappa) = \{\mathbf{x} \in \mathbb{R}^n : \kappa \mathbf{x} \in \boldsymbol{\mathcal{D}}\}.$$

Then $\operatorname{diam}(\boldsymbol{\mathcal{D}}(\kappa)) = \frac{1}{\kappa} \operatorname{diam}(\boldsymbol{\mathcal{D}})$. Let $\mu \equiv 1$ so that $\operatorname{curl} \mathbf{a}_N \equiv \mathbf{e}$ in $\boldsymbol{\mathcal{D}}(\kappa)$. Consider (ψ, A) satisfying (1.2) on $\boldsymbol{\mathcal{D}}(\kappa)$. Let

$$\widetilde{\psi}(\mathbf{x}) = \psi(\kappa^{-1}\mathbf{x}) \quad \text{ for } \quad \mathbf{x} \in \boldsymbol{\mathcal{D}}$$

and

$$\widetilde{\mathbf{A}}(\mathbf{x}) = \mathbf{A}(\kappa^{-1}\mathbf{x}) \quad \text{for} \quad \mathbf{x} \in \mathbb{R}^3.$$

Then $(\widetilde{\psi}, \widetilde{\mathbf{A}})$ satisfies

$$\text{(4.2)} \qquad \begin{cases} (i\nabla + \widetilde{\mathbf{A}})^2 \widetilde{\psi} - \widetilde{\psi} + |\widetilde{\psi}|^2 \widetilde{\psi} = 0 & \text{in } \boldsymbol{\mathcal{D}}, \\ \operatorname{curl}^2 \widetilde{\mathbf{A}} = -\kappa^{-2} \left(\frac{i}{2} (\widetilde{\psi}^* \nabla \widetilde{\psi} - \widetilde{\psi} \nabla \widetilde{\psi}^*) + \widetilde{\mathbf{A}} |\psi|^2 \right) \chi_{\boldsymbol{\mathcal{D}}} & \text{in } \mathbb{R}^3, \\ n \cdot (i\nabla + \widetilde{\mathbf{A}}) \widetilde{\psi} = -i\gamma \widetilde{\psi} & \text{on } \partial\boldsymbol{\mathcal{D}}, \\ (\operatorname{curl} \widetilde{\mathbf{A}} - h\kappa^{-1}\mathbf{e}) \in L^2(\mathbb{R}^3; \mathbb{R}^3). \end{cases}$$

Assume that $0 < \kappa \leq 1$. Arguing as in Lemma 2.5, taking into account the multiple κ^{-2} of the right-hand side of the second equation in (4.2), we obtain the analogue of (2.7),

$$\int_{\boldsymbol{\mathcal{D}}} |\widetilde{\mathbf{A}} - h\kappa^{-1}\mathbf{a}_N|^2 d\mathbf{x} \leq \kappa^{-4} M \int_{\boldsymbol{\mathcal{D}}} |\widetilde{\psi}|^2 d\mathbf{x},$$

where $M = M(\operatorname{diam} \boldsymbol{\mathcal{D}})$. Using $\kappa \leq 1$ we then obtain the analogue of (2.5),

$$\int_{\boldsymbol{\mathcal{D}}} |(i\nabla + h\kappa^{-1}\mathbf{a}_N)\widetilde{\psi}|^2 d\mathbf{x} \leq C_1 \kappa^{-4} \int_{\boldsymbol{\mathcal{D}}} |\widetilde{\psi}|^2 d\mathbf{x}.$$

It follows as in Theorem 2.9 then that there is a constant ϕ so that if $h\kappa^{-1} \geq \phi\kappa^{-4}$ then $\widetilde{\psi} \equiv 0$. Returning to $(\psi, \mathbf{A})$ we conclude that

$$\overline{h}(\kappa, 1, \gamma, \boldsymbol{\mathcal{D}}(\kappa)) = O(\kappa^{-3}) \qquad \text{as } \kappa \downarrow 0.$$

As we alluded to in the introduction, there are superconducting solutions for the case of the slab $-\infty < x < 0$, $-\infty < y, z < \infty$ for which h diverges as $\kappa \downarrow 0$ (see [4]). However, this occurs at the slower rate $h = O(\kappa^{-\frac{1}{2}})$.

REFERENCES

[1] P. Bauman, D. Phillips, and Q. Tang, *Stable nucleation for the Ginzburg–Landau system with an applied magnetic field*, Arch. Rationál Mech. Anal., 142 (1998), pp. 1–43.

[2] C. Bolley, *Modélization de champ de retard à la condensation d'un supraconducteur par un problème de bifurcation*, M^2AN, 26 (1992), pp. 235–287.

[3] C. Bolley and B. Helffer, *An application of semi-classical analysis to the asymptotic study of the supercooling field of a superconducting material*, Ann. Inst. H. Poincaré, 58 (1993), pp. 89–233.

[4] C. Bolley and B. Helffer, *On the asymptotics of the critical fields for the Ginzburg–Landau equation*, Progress in Partial Differential Equations: the Metz Surveys, Pitman Res. Notes Math. Ser. 314, 1994, pp. 18–32, Longman Scientific and Technical, Harlow.

[5] J. Chapman, Q. Du, and M. Gunzburger, *A Ginzburg–Landau type model for superconducting/normal junctions including Josephson junctions*, European J. Appl. Math., 6 (1995), pp. 97–114.

[6] M. Dauge and B. Helffer, *Eigenvalues variation,* I*, Neumann problem for Sturm–Liouville operators*, J. Differential Equations, 104 (1993), pp. 243–262.

[7] Q. Du, M. D. Gunzburger, and J. Peterson, *Analysis and approximation of the Ginzburg–Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[8] P. G. de Gennes, *Superconductivity: Selected topics in solid state physics and theoretical physics*, in Proc. 8th Latin American School of Physics, Caracas, 1966.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, 1977.

[10] O. A. Ladyzhenskaya, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York 1969.

[11] O. A. Ladyzhenskaya, *The Boundary Value Problems of Mathematical Physics*, Springer-Verlag, Berlin, New York, 1985.

[12] The Orsay Group, *Strong field effects at the surface of a superconductor*, in Quantum Fluids, D. F. Bewer, ed., North–Holland, Amsterdam, 1966, p. 26.

[13] D. Saint-James and P. G. de Gennes, *Onset of superconductivity in decreasing fields*, Phys. Lett., 7 (1963), pp. 306–308.

[14] M. Tinkham, *Introduction to Superconductivity*, 2nd ed., McGraw–Hill, New York, 1996.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 360–378

GLOBAL STABILITY AND PERMANENCE FOR A CLASS OF TYPE *K* MONOTONE SYSTEMS*

TU CAIFENG† AND JIANG JIFA†

Abstract. In this paper, the convergence of solutions of type K monotone systems is studied. The basic assumption is that the Jacobian matrix is stable for every point in R^n_+. The main results are the following. If the system has a positive steady state, then it is globally asymptotically stable in $\mathrm{Int}R^n_+$. A sufficient and necessary condition for a nonnegative steady state of the system to be globally asymptotically stable is presented. Moreover, we provide sufficient conditions for type K monotone systems to be permanent and for existence and uniqueness of a positive steady state.

Key words. type K monotone system, steady state, global stability, permanence

AMS subject classifications. Primary, 34D20; Secondary, 90A16

PII. S0036141097325290

1. Introduction.

Beginning with the ground-breaking work of M. W. Hirsch [1], [2], [3] for cooperative systems and monotone semiflows, this direction has recently received considerable attention. H. L. Smith [4] has given a wonderful overview of the current state of the theory of monotone semiflows and has nicely illustrated the theory with applications to systems of ordinary and delay differential equations and reaction-diffusion equations. Hirsch [1], [2], [3] not only proved a series of important results on the long-run behavior of trajectories for monotone semiflows but also presented many powerful key ideas to deal with the asymptotic behavior of solutions of many sorts of differential equations preserving some type of order relation on initial data, boundary data, and inhomogeneous terms. As soon as one proves that a system preserves a suitable order relation on the state space, he can either apply the abstract results for monotone semiflows to such a system or directly study it by using the ideas presented by Hirsch.

In the paper [5], Smith studied general Kolmogorov-type models of competition between subcommunities which are described by the following ordinary differential equations:

$$\dot{x}_i = x_i f_i(x_1, x_2, \ldots, x_n), \quad 1 \le i \le n, \quad x_i \ge 0. \tag{S}$$

More precisely, he considered systems (S) for which, after suitable permutations of species indices, the community consists of two disjoint complementary subcommunities $I = \{1, 2, \ldots, k\}$ and $J = \{k+1, \ldots, n\}, 0 \le k \le n$. The interactions between every pair of species in subcommunity I are mutualistic and similarly for group J. On the other hand, the interaction between species $i \in I$ and species $j \in J$ is a

*Received by the editors July 29, 1997; accepted for publication January 13, 1998; published electronically January 5, 1999. This research was supported by the National Natural Science Foundation of China.

http://www.siam.org/journals/sima/30-2/32529.html

†Department of Mathematics, University of Science and Technology of China, Hefei, P.R. China (rengb@ustc.edu.cn).

competitive one. Mathematically speaking, the above situation can be expressed as

$$\text{(KM)} \qquad \begin{cases} \dfrac{\partial f_i}{\partial x_j} \geq 0 \quad \text{whenever} \quad i \neq j \quad \text{belongs to the same subcommunity,} \\ \dfrac{\partial f_i}{\partial x_j} \leq 0 \quad \text{whenever} \quad i, \quad j \quad \text{belong to different subcommunities.} \end{cases}$$

Smith called such a system (S) a type monotone system.

Earlier Takeuchi and Adachi [6], Takeuchi, Adachi, and Tokumaru [7], and Travis and Post [8] had investigated some particular sorts of such systems, most of which are Lotka–Volterra systems

$$\text{(S}^*\text{)} \qquad \dot{x} = \operatorname{diag}(x)(r + Mx), \quad x \in R_+^n, \quad r \in R^n,$$

in which $x = (x_1, x_2) \in R_+^k \times R_+^{n-k}$, $r = (r_1, r_2)$, and

$$M = \begin{pmatrix} A & -B \\ -C & D \end{pmatrix}, \tag{1.1}$$

where A is a $k \times k$ matrix with nonnegative off-diagonal elements, D is an $(n-k)\times(n-k)$ matrix with the same property, and $B \geq 0, C \geq 0$. Remarkable results for systems (S*) had been obtained by them in [6], [7], [8] without using the monotonicity properties of the semiflows. Some of these results rely on Lyapunov function arguments due to Goh [9].

Smith [5] first discovered that these systems preserve the partial ordering generated by the cone $K = R_+^k \times (-R_+^{n-k})$ and exploited the monotonicity of the semiflows to study such more general Kolmogorov models. In that paper, he extended Hirsch's ideas for cooperative systems to type K monotone systems (S) and found sufficient conditions for persistence of all species. In his analysis, two subsystems of (S) play a major role. These subsystems model the two mutualistic subcommunities I and J in isolation:

$$\text{(S}_\text{I}\text{)} \qquad \dot{x}_i = x_i f_i(x_1, \ldots, x_k, 0, \ldots, 0), \quad x_i \geq 0, \quad i \in I,$$

and

$$\text{(S}_\text{J}\text{)} \qquad \dot{x}_j = x_j f_j(0, \ldots, 0, x_{k+1}, \ldots, x_n), \quad x_j \geq 0, \quad j \in J.$$

He accomplished his persistent results in the case that (S$_\text{I}$) and (S$_\text{J}$) possess positive steady states and that each of these positive steady states can be successfully invaded by each of the competing species. These results are valid only for nonobligate species (that is, $f(0) > 0$). We generalized them to obligate cases in [10].

Takeuchi and Adachi [6] made use of results from mathematical programming and Lyapunov function arguments to type K Lotka–Volterra systems (S*) and obtained more sharp results. They proved that if M is stable, then (S*) has a unique nonnegative steady state x^* with the property that x^* attracts all solutions with positive initial conditions. Indeed, if $L = \{i : x_i^* > 0\}$, then the domain of attraction of x^* is $\{x \in R_+^n : x_i > 0 \quad \text{for all} \quad i \in L\}$. They also proved that if M is stable and x^* is a nonnegative steady state with $x_i^* > 0$ for $i \in L$ and $x_i^* = 0$ for $i \in N/L$ where $N = \{1, 2, \ldots, n\}$, then x^* is globally asymptotically stable relative to

$$\{x \in R_+^n : x_i > 0 \quad \text{for all} \quad i \in L\}$$

if and only if

$$r_p + \sum_{j=1}^{n} m_{p_j} x^*{}_j \le 0 \quad \text{for all} \quad p \in N/L, \tag{1.2}$$

which is equivalent to

$$r + Mx^* \le 0. \tag{1.3}$$

As a simple application of these results, one can easily obtain the following two propositions.

(P_1). If (S_I^*) and (S_J^*) have positive steady states x_1^0 and x_2^0, which are globally asymptotically stable in $\mathrm{Int}R_+^k$ and $\mathrm{Int}R_+^{n-k}$, respectively, and each is stable to invasion by every competing species,

$$r_2 - Cx_1^0 \le 0$$

and

$$r_1 - Bx_2^0 \le 0,$$

then (S^*) cannot have a stable positive steady state.

(P_2). If M is stable and x_2^0 is a positive steady state of (S_J^*) which cannot be invaded by species $i \in I$,

$$r_1 - Bx_2^0 \le 0,$$

then $(0, x_2^0)$ is globally asymptotically stable with respect to $\{x \in R_+^n : x_j > 0, \quad j \in J\}$.

The proposition (P_1) shows that if the two positive steady states of (S_I^*) and (S_J^*) representing stable persistence of each subcommunity I and J in the absence of competition are such that each is stable to invasion by every competing species, then there cannot be a stable positive steady state representing coexistence of the two subgroups. The proposition (P_2) shows that if a positive steady state of (S_J^*) is stable to invasion by each species $i \in I$, then all solutions approach that steady state; the species $i \in I$ die out.

Based on the above conclusions for Lotka–Volterra systems, Smith [5, p. 870] pointed out the following two questions:

"Are there sufficient conditions, symmetrical with respect to the two subgroups I and J, for the conclusion (P_1) to hold for the more general type K monotone systems (S)?" and

"Are there sufficient conditions, which need not be symmetrical with respect to I and J, for which the proposition (P_2) generalizes to type K monotone systems (S)?"

The main goal of this paper is to introduce suitable conditions to answer the above two questions. The first condition we shall introduce is

$$Df(x) \le_K M \quad \text{for every } x \in R_+^n, \tag{C_1}$$

where the constant matrix M is of type K and stable. Under this condition, we shall completely generalize those results of Takeuchi and Adachi [6] to more general type K monotone systems (S). More precisely, we shall prove that if type K monotone systems (S) satisfy (C_1), then (S) has a unique steady state $c \in R_+^n$ whose attraction

domain is $\Omega = \{x \in R_+^n : x_i > 0 \quad \text{for} \quad i \in L = \{i : c_i > 0\}\}$, and that a steady state c is globally asymptotically stable relative to Ω if and only if $f(c) \leq 0$. These results obviously give an answer to each of the above two questions.

The second condition we shall present is the concave one:

$$(C_2) \qquad Df(y) \leq_K Df(x) \quad \text{whenever} \quad 0 \leq x \leq_K y.$$

We shall verify that if type K monotone systems (S) satisfy (C_2), then the same conclusion as (P_1) holds. This result gives another answer to the first question. Moreover, we shall provide sufficient conditions for systems (S) to be permanent and for existence and uniqueness of a positive steady state.

2. Preliminary. In this section, we will introduce some notation, establish some conventions, and describe some results which are essential tools in the later sections.

Let $R_+^n = \{x \in R^n : x_i \geq 0 \quad \text{for} \quad 1 \leq i \leq n\}$ denote the nonnegative orthant and $\text{Int}R_+^n = \{x \in R_+^n : x_i > 0 \quad \text{for} \quad 1 \leq i \leq n\}$ denote its interior.

The idea of a cone in R^n and the associated partial ordering which the cone generates will be of fundamental importance in this paper. Recall that a cone K in R^n is a closed convex set of R^n with the property $K \cap (-K) = \{0\}$. It is easy to see that R_+^n and $K = \{x \in R^n : x_i \geq 0 \quad \text{for} \quad 1 \leq i \leq k \quad \text{and} \quad x_j \leq 0 \quad \text{for} \quad k+1 \leq j \leq n\}$ are two cones in R^n. We write $x \leq_K y$ $(x \leq y)$ whenever $y - x \in K$ $(y - x \in R_+^n)$ and $x <_K y$ $(x < y)$ whenever $y - x \in \text{Int}K$ $(y - x \in \text{Int}R_+^n)$. If $x, y \in R^n$ and K is a cone in R^n, we let $[x, y]_K = \{z \in R^n : x \leq_K z \leq_K y\}$.

If A is an $n \times m$ matrix, we write $A \geq 0$ if $a_{ij} \geq 0$ for all i and j. If M is an $n \times n$ matrix and $M(R_+^k \times (-R_+^{n-k})) \subset R_+^k \times (-R_+^{n-k})$, then it is easy to see that M has structure (1.1) where $A, B, C, D \geq 0$. If this is the case, we write $M \geq_K 0$ where $K = R_+^k \times (-R_+^{n-k})$. A matrix having structure (1.1) is called a type K matrix. For two $n \times n$ type K matrices M_1 and M_2, $M_1 \geq_K M_2$ if and only if $M_1 - M_2 \geq_K 0$; that is, $A_1 \geq A_2$, $B_1 \geq B_2$, $C_1 \geq C_2$, $D_1 \geq D_2$.

We will reserve the letter n for the dimension of space R^n and $N = \{1, 2, \ldots, n\}$. Let L be a nonempty subset of N and $\bar{L} = N/L$ be its complementary set in N. Then the set $H_L^+ = \{x \in R_+^n : x_p = 0 \quad \text{for} \quad p \in \bar{L}\}$ is an invariant set for (S). The $H_L^+, L \subset N$, make up the boundary of R_+^n. If $L \subset N$, $L = \{i_1 < \cdots < i_{r_1} < j_1 < \cdots < j_{r_2}\}$ for $i_h \in I, 1 \leq h \leq r_1$ and $j_m \in J, 1 \leq m \leq r_2$, then for $x \in R_+^n$ we write $x_L = \{x_{i_1}, \ldots, x_{i_{r_1}}, x_{j_1}, \ldots, x_{j_{r_2}}\} \in R_+^l$, where $r_1 + r_2 = l = \#L$. For any $x \in H_L^+$, we write $x = (x_L, 0)$ and $f_L(x_L, 0) = (f_{i_1}(x_L, 0), \ldots, f_{i_{r_1}}(x_L, 0), f_{j_1}(x_L, 0), \ldots, f_{j_{r_2}}(x_L, 0))$. Then the subsystem of (S) obtained by setting $x_p = 0$ for $p \in \bar{L}$ will be denoted by

$$(S_L) \qquad \dot{x}_L = \text{diag}(x_L) f_L(x_L, 0), \quad x_L \in R_+^l,$$

and the dynamics on H_L^+ is determined by (S_L). For systems (S) and (S_L), we make corresponding Lotka–Volterra systems (S^*) and (S_L^*) which will play an important role in our analysis:

$$(S^*) \qquad \dot{x} = \text{diag}(x)(r + Mx), \quad x \in R_+^n, \quad r \in R^n,$$

and

$$(S_L^*) \qquad \dot{x}_L = \text{diag}(x_L)(r_L + M_L x_L), \qquad x_L \in R_+^l, \quad r_L \in R^l,$$

where M is a stable type K matrix such that the condition (C_1) is satisfied. The meaning of the inequality $Df(x_L, 0) \leq_K M$ implies that $Df_L(x_L, 0) \leq_{K_L} M_L$, where

$Df_L(x_L,0)$ is an $l\times l$ submatrix of $Df(x_L,0)$ obtained by deleting rows and columns of $Df(x_L,0)$ indexed by $p\in\bar{L}$, M_L is an $l\times l$ submatrix of M obtained in the same way, and $K_L=R_+^{r_1}\times(-R_+^{r_2})$. Later $Df_L(x_L,0)$ and M_L will have the same meaning.

We let $s(M)=\max\ \mathrm{Re}\,\lambda$, where λ runs through the eigenvalues of M. M is stable if $s(M)<0$. From the Perron–Frobenius theory, $s(M)$ is an eigenvalue of M if M is of type K.

We write $\phi_t(x)$, $\psi_t(x)$ for the unique solutions $x(t)$ of (S) and $\widetilde{x}(t)$ of (S*), respectively, satisfying $x(0)=x$, $\widetilde{x}(0)=x$. Similarly, $\phi_t^L(x_L)$, $\psi_t^L(x_L)$ are the unique solutions $x_L(t)$ of ($\mathrm{S_L}$) and $\widetilde{x}_L(t)$ of ($\mathrm{S_L^*}$), respectively, satisfying $x_L(0)=x_L$ and $\widetilde{x}_L(0)=x_L$. $\{\phi_t(x)\}_i$ for $i\in N$ represents the ith component of $\phi_t(x)$. The components $\{\psi_t(x)\}_i$, $\{\phi_t^L(x_L)\}_i$, $\{\psi_t^L(x_L)\}_i$ for $i\in N$ are defined similarly.

One of the fundamental tools used in this paper is the generalized Kamke theorem. This theorem is extended in a natural way from cooperative systems to type K monotone systems (see [5, Theorem 2.4]).

THEOREM 2.1 (Kamke theorem). *Assume that the system* (S) *is of type K monotone, and $x(t),y(t)$ are the solutions of* (S) *defined on $a\le t\le b$ with $x(a)\le_K y(a)$. Then $x(t)\le_K y(t)$ for all $t\in[a,b]$.*

In the paper [5, p. 862], Smith developed a criterion for the monotonicity of every component of a solution for a cooperative system given by Selgrade [11] to the type K monotone system. Now we quote it here.

THEOREM 2.2. *Let the system* (S) *be a type K monotone system and let $f(x)\ge_K 0$ for some $x\in R_+^n$. Then $\{\phi_t(x)\}_i$ is nondecreasing if $i\in I$ and $\{\phi_t(x)\}_j$ is nonincreasing if $j\in J$ for all $t\ge 0$ for which the solution exists. A similar result holds if $f(x)\le_K 0$.*

The following result is from Perron–Frobenius theory (see [5, pp. 861 and 873]).

THEOREM 2.3. *Let M, N be of type K, where $K=R_+^k\times(-R_+^{n-k})$.*

(i) *If $N\ge_K M$, then $s(N)\ge s(M)$.*

(ii) *If $s(M)<0$, then $s(M_L)<0$ for any $L\subset N$.*

(iii) *If $s(M)<0$, then $(-M)^{-1}\ge_K 0$.*

3. The global stability for a positive steady state. The object of this section is to prove the following theorem.

THEOREM 3.1. *Let the conditions* (KM) *and* ($\mathrm{C_1}$) *hold. If* (S) *has a positive steady state $\bar{x}$, then $\bar{x}$ is globally asymptotically stable in* $\mathrm{Int}R_+^n$.

Proof. Let $\bar{x}$ be a positive steady state of (S), that is, $f(\bar{x})=0$. Then for any x in R_+^n, we have

$$f(x)-f(\bar{x})=\left[\int_0^1 Df(sx+(1-s)\bar{x})ds\right](x-\bar{x}).$$

The condition ($\mathrm{C_1}$) shows that

$$\int_0^1 Df(sx+(1-s)\bar{x})ds\le_K M. \tag{3.1}$$

Then the meaning of (3.1) implies that

$$f(x)\le_K M(x-\bar{x})\quad\text{for}\quad x\ge_K\bar{x} \tag{3.2}$$

and

$$f(x)\ge_K M(x-\bar{x})\quad\text{for}\quad x\le_K\bar{x}. \tag{3.3}$$

From (3.2) and (3.3), we obtain that

$$\operatorname{diag}(x)f(x) \leq_K \operatorname{diag}(x)(r + Mx) \quad \text{for} \quad x \geq_K \bar{x}$$

and

$$\operatorname{diag}(x)f(x) \geq_K \operatorname{diag}(x)(r + Mx) \quad \text{for} \quad x \leq_K \bar{x},$$

where $r = -M\bar{x}$.

Consider the systems (S) and (S*). It follows from standard differential inequality arguments and Theorem 2.1 that

$$\bar{x} \leq_K \phi_t(x) \leq_K \psi_t(x) \quad \text{for} \quad x \geq_K \bar{x}, \quad t \geq 0, \tag{3.4}$$

and

$$\bar{x} \geq_K \phi_t(x) \geq_K \psi_t(x) \quad \text{for} \quad x \leq_K \bar{x}, \quad t \geq 0. \tag{3.5}$$

Since $M\bar{x} + r = 0$, $\bar{x}$ is a positive steady state of (S*). Because $s(M) < 0$, it follows that

$$\lim_{t\to+\infty} \psi_t(x) = \bar{x} \quad \text{for any} \quad x \in \operatorname{Int}R^n_+$$

from the results of Takeuchi and Adachi [6] mentioned in the introduction. Then it deduces that

$$\lim_{t\to+\infty} \phi_t(x) = \bar{x} \quad \text{for} \quad x \in \operatorname{Int}R^n_+ \quad \text{with} \quad x \geq_K \bar{x} \quad \text{or} \quad x \leq_K \bar{x}$$

from (3.4) and (3.5).

For any $x \in \operatorname{Int}R^n_+$, there exist $y \in (\bar{x} + K) \cap \operatorname{Int}R^n_+$ and $z \in (\bar{x} - K) \cap \operatorname{Int}R^n_+$ such that $z \leq_K x \leq_K y$. In fact, we choose y and z as follows:

$$\begin{cases} y_i = \max(x_i, \bar{x}_i) & \text{for} \quad i \in I, \\ y_j = \min(x_j, \bar{x}_j) & \text{for} \quad j \in J, \end{cases} \tag{3.6}$$

and

$$\begin{cases} z_i = \min(x_i, \bar{x}_i) & \text{for} \quad i \in I, \\ z_j = \max(x_j, \bar{x}_j) & \text{for} \quad j \in J. \end{cases} \tag{3.7}$$

Applying Theorem 2.1, we have

$$\phi_t(z) \leq_K \phi_t(x) \leq_K \phi_t(y) \quad \text{for} \quad t \geq 0.$$

Because

$$\lim_{t\to+\infty} \phi_t(z) = \lim_{t\to+\infty} \phi_t(y) = \bar{x},$$

we have established that

$$\lim_{t\to+\infty} \phi_t(x) = \bar{x} \quad \text{for all} \quad x \in \operatorname{Int}R^n_+.$$

This completes the proof. □

4. The global stability for a saturated steady state. For convenience, we denote the set of steady states for (S) by E. A steady state $c \in E$ satisfying that $\operatorname{diag}(c)f(c) = 0$ and $f(c) \leq 0$ is called a saturated steady state.

THEOREM 4.1. *Let the assumption* (KM) *and the condition* (C_1) *hold. If there exists a steady state* $c \in \partial R^n_+$ *with* $c_P > 0$, $c_{\bar{P}} = 0$ *where we agree on* $c = 0$ *if* $P = \phi$, *then* c *attracts all solutions with initial conditions in* $\{x \in R^n_+ : x_P > 0\}$ *if and only if* $f(c) \leq 0$.

Before proving Theorem 4.1, we again stress some conventions which will be employed throughout the later work.

If $x \in R^n_+$, let $x_L = \{x_{i_1}, \ldots, x_{i_{r_1}}, x_{j_1}, \ldots, x_{j_{r_2}}\} \in R^l_+$, where $r_1 + r_2 = l = \#L$, $i_h \in I$, $1 \leq h \leq r_1$ and $j_m \in J$, $1 \leq m \leq r_2$. If $x \in H^+_L$, we write $x = (x_L, 0)$ and $f_L(x_L, 0) = (f_{i_1}(x_L, 0), \ldots, f_{i_{r_1}}(x_L, 0), f_{j_1}(x_L, 0), \ldots, f_{j_{r_2}}(x_L, 0))$.

In the proof of Theorem 4.1, we will use the following proposition, which can be found in [5, p. 864].

PROPOSITION 4.2. *Let* $x = (x_I, x_J) \in R^n_+$; *then* $\phi_t(x) \leq (\phi^I_t(x_I), \phi^J_t(x_J))$ *for* $t \geq 0$.

Proof of Theorem 4.1 (sufficiency). Assume that $P = \phi$; that is, $c = 0$ where ϕ is an empty set. In this case, $f(0) \leq 0$. Considering the subsystems (S_I) and cooperative system (S^*_I)

$$\dot{x}_I = \operatorname{diag}(x_I)[f_I(0) + M_I x_I],$$

we have

$$f_I(x_I, 0) - f_I(0) = \left[\int_0^1 Df_I(sx_I, 0)ds\right] x_I \leq M_I x_I.$$

Then

$$\operatorname{diag}(x_I)f_I(x_I, 0) \leq \operatorname{diag}(x_I)[f_I(0) + M_I x_I],$$

and

$$0 \leq \phi^I_t(x_I) \leq \psi^I_t(x_I) \quad \text{for} \quad t \geq 0.$$

Since $s(M_I) < s(M) < 0$ and $f_I(0) \leq 0$, by a result of Jiang (see [12, Theorem 3.4]), $\lim_{t\to+\infty} \psi^I_t(x_I) = 0_I$ holds. So $\lim_{t\to+\infty} \phi^I_t(x_I) = 0_I$. Similarly, we obtain that $\lim_{t\to+\infty} \phi^J_t(x_J) = 0_J$. Hence, $\lim_{t\to+\infty} \phi_t(x) = 0$ by Proposition 4.2.

If $P \neq \phi$, it is assumed that there exists a steady state $c \in \partial R^n_+$ with $c_P > 0$, $c_{\bar{P}} = 0$ for $P \subset N$, and $f(c) \leq 0$. Then we divide the proof into three cases.

$$\text{(i)} P \subset I, \qquad \text{(ii)} P \subset J, \qquad \text{(iii)} P \cap I \neq \phi, \quad P \cap J \neq \phi.$$

Case (i). Without loss of generality, we may assume that $P = \{1, 2, \ldots, p\}$, $p = \#P$, $1 \leq p \leq k$. Then $c = (c_1, \ldots, c_p, 0, \ldots, 0) = (c_P, 0)$.

Define a set

$$\Omega_1 = \{x \in R^n_+ : \quad x_i \geq c_i \quad \text{for} \quad i = 1, \ldots, k, \quad x_j = 0 \quad \text{for} \quad j = k+1, \ldots, n\}.$$

Consider the subsystem (S_I)

$$\dot{x}_I = \operatorname{diag}(x_I)f_I(x_I, 0), \quad x_I \in R^k_+.$$

For any $x_I \in R_+^k$, we have

$$f_I(x_I, 0) - f_I(c_I, 0) = \left[\int_0^1 Df_I(sx_I + (1-s)c_I, 0)ds\right](x_I - c_I).$$

The condition (C_1) implies that

$$\int_0^1 Df_I(sx_I + (1-s)c_I, 0)ds \le M_I.$$

Then

$$f_I(x_I, 0) \le f_I(c_I, 0) - M_I c_I + M_I x_I \quad \text{for} \quad x_I \ge c_I.$$

It is easy to obtain that

$$\text{diag}(x_I)f_I(x_I, 0) \le \text{diag}(x_I)(f_I(c_I, 0) - M_I c_I + M_I x_I) \quad \text{for} \quad x_I \ge c_I. \tag{4.1}$$

Make a corresponding cooperative system (S_I^*)

$$\dot{x}_I = \text{diag}(x_I)(f_I(c_I, 0) - M_I c_I + M_I x_I)$$

and write $\psi_t^I(x_I)$ for the unique solution $\widetilde{x}_I(t)$ of (S_I^*) satisfying $\widetilde{x}_I(0) = x_I$.

Obviously, $x_I = c_I$ is a steady state of (S_I^*) and

$$[f_I(c_I, 0) - M_I c_I + M_I x_I]_{x_I = c_I} = f_I(c_I, 0) \le 0.$$

Thus, Theorem 3.4 in [12] implies that

$$\lim_{t \to +\infty} \psi_t^I(x_I) = c_I \quad \text{for any} \quad x_I \in R_+^k \quad \text{with} \quad x_P > 0.$$

From (4.1), by standard differential inequality arguments, we have

$$c_I \le \phi_t^I(x_I) \le \psi_t^I(x_I) \quad \text{for} \quad x_I \ge c_I \quad \text{and} \quad t \ge 0.$$

Thus

$$\lim_{t \to +\infty} \phi_t^I(x_I) = c_I$$

for any $x_I \in R_+^k$ with $x_I \ge c_I$.

Observe that if $x \in \Omega_1$, then $x \ge_K c$ with $x_P > 0$. Because $\{\phi_t(x_I, 0)\}_j = 0$ for $j \in J$, we write $\phi_t(x) = (\phi_t^I(x_I), 0)$. Thus

$$\lim_{t \to +\infty} \phi_t(x) = \lim_{t \to +\infty} (\phi_t^I(x), 0) = (c_I, 0) = c$$

for $x \in \Omega_1$.

Define another set

$$\begin{aligned}\Omega_2 = \{x \in R_+^n : \quad & 0 < x_i \le c_i \quad \text{for} \quad i = 1, \dots, p, \quad x_i = 0 \\ & \text{for} \quad i = p+1, \dots, k, \quad x_j \ge 0 \quad \text{for} \quad j = k+1, \dots, n\}.\end{aligned}$$

Let $L = \{1, \dots, p, k+1, \dots, n\}$, $\#L = n - k + p$. We consider the subsystem (S_L)

$$\dot{x}_L = \text{diag}(x_L)f_L(x_L, 0), \quad x_L \in R_+^{n-k+p}.$$

For any $x_L \in R_+^{n-k+p}$, we have

$$f_L(x_L,0) - f_L(c_L,0) = \left[\int_0^1 Df_L(sx_L + (1-s)c_L, 0)ds\right](x_L - c_L). \tag{4.2}$$

If $x \in \Omega_2$, then $x_L \leq_{K_1} c_L$ with $x_P > 0$ where $K_1 = R_+^p \times (-R_+^{n-k})$. Since

$$\int_0^1 Df_L(sx_L + (1-s)c_L, 0)ds \leq_{K_1} M_L,$$

(4.2) shows that

$$f_L(x_L,0) \geq_{K_1} f_L(c_L,0) + M_L(x_L - c_L) = r_L + M_L x_L,$$

where $r_L = f_L(c_L,0) - M_L c_L$. It follows that

$$\operatorname{diag}(x_L) f_L(x_L,0) \geq_{K_1} \operatorname{diag}(x_L)(r_L + M_L x_L) \quad \text{for} \quad x_L \leq_{K_1} c_L. \tag{4.3}$$

Constructing a corresponding Lotka–Volterra system (S_L^*), we have that

$$\dot{x}_L = \operatorname{diag}(x_L)(r_L + M_L x_L), \quad x_L \in R_+^{n-k+p}, \quad r_L \in R^{n-k+p},$$

where $r_L = f_L(c_L,0) - M_L c_L$. Evidently,

$$\operatorname{diag}(c_L)(r_L + M_L c_L) = \operatorname{diag}(c_L) f_L(c_L,0) = 0.$$

This means that c_L is a steady state of (S_L^*) and

$$r_L + M_L c_L = f_L(c_L,0).$$

By assumption, $f_L(c_L,0) \leq 0$ is true. Then according to the results of Takeuchi and Adachi [6], we deduce that

$$\lim_{t\to+\infty} \psi_t^L(x_L) = c_L$$

for any $x_L \in \{u \in R_+^{n-k+p} : u_j > 0 \quad \text{for} \quad j \in P\}$. From (4.3), by standard differential inequality arguments, we have

$$c_L \geq_{K_1} \phi_t^L(x_L) \geq_{K_1} \psi_t^L(x_L) \quad \text{for} \quad t \geq 0.$$

Then

$$\lim_{t\to+\infty} \phi_t^L(x_L) = c_L$$

for any $x_L \in \{u \in R_+^{n-k+p} : u_j > 0 \quad \text{for} \quad j \in P\}$.

If $x \in \Omega_2$ which implies that $x \leq_K c$ with $x_i > 0$ for $i \in P$, then

$$\phi_t(x) = (\{\phi_t(x)\}_1, \ldots, \{\phi_t(x)\}_p, 0, \ldots, 0, \{\phi_t(x)\}_{k+1}, \ldots, \{\phi_t(x)\}_n) = (\phi_t^L(x_L), 0).$$

We have established that

$$\lim_{t\to+\infty} \phi_t(x) = \lim_{t\to+\infty} (\phi_t^L(x_L), 0) = (c_L, 0) = c \quad \text{for any} \quad x \in \Omega_2.$$

Hence, c attracts all points in Ω_1 and Ω_2.

For any $x \in R_+^n$ with $x_i > 0$ for $i \in P$, there exist $y \in \Omega_1$ and $z \in \Omega_2$ such that

$$z \leq_K x \leq_K y.$$

Indeed, we can choose y and z in a similar way as (3.6) and (3.7) if $\bar{x}$ is replaced by c. Theorem 2.1 shows that

$$\phi_t(z) \leq_K \phi_t(x) \leq_K \phi_t(y) \quad \text{for} \quad t \geq 0.$$

Since

$$\lim_{t\to+\infty} \phi_t(z) = \lim_{t\to+\infty} \phi_t(y) = c,$$

we conclude that

$$\lim_{t\to+\infty} \phi_t(x) = c.$$

Case (ii). If Case (ii) is true, then pick a transformation by $(x_1, \ldots, x_n)$ such that it is changed into Case (i). Actually, we only choose a suitable permutations of species indices for this purpose.

Case (iii). Without loss of generality, let

$$c = (c_1, \ldots, c_{r_1}, 0, \ldots, 0, c_{k+1}, \ldots, c_{k+r_2}, 0, \ldots, 0).$$

Then $P = \{1, \ldots, r_1, k+1, \ldots, k+r_2\}$, $1 \leq r_1 \leq k$, $1 \leq r_2 \leq n-k$, $r_1 + r_2 = p < n$.

Define a set

$$\begin{aligned}\Omega_3 = \{x \in R_+^n : x_i \geq c_i \quad &\text{for} \quad i = 1, \ldots, k, \quad 0 < x_j \leq c_j \quad \text{for} \quad j = k+1, \ldots, k+r_2,\\ &x_j = 0 \quad \text{for} \quad j = k+r_2+1, \ldots, n\}\\ = (c + K) \cap R_+^n \cap \{x \in R_+^n : x_P &> 0\}.\end{aligned}$$

Consider the systems $(\mathrm{S_L})$ and $(\mathrm{S_L^*})$ with $L = \{1, \ldots, k, k+1, \ldots, k+r_2\}$.

If $x \in \Omega_3$, then $x_L \geq_{K_2} c_L$ where $K_2 = R_+^k \times (-R_+^{r_2})$. Letting $r_L = f_L(c_L, 0) - M_L c_L$, similar to Case (i) we can obtain that

$$\operatorname{diag}(x_L) f_L(x_L, 0) \leq_{K_2} \operatorname{diag}(x_L)(r_L + M_L x_L), \quad x_L \in R_+^{k+r_2} \quad \text{with} \quad x_L \geq_{K_2} c_L,$$

and

$$c_L \leq_{K_2} \phi_t^L(x_L) \leq_{K_2} \psi_t^L(x_L) \quad \text{for} \quad x_L \geq_{K_2} c_L \quad \text{and} \quad t \geq 0.$$

We may argue as in Case (i) with the help of the results of Takeuchi and Adachi [6]. It deduces that

$$\lim_{t\to+\infty} \phi_t^L(x_L) = \lim_{t\to+\infty} \psi_t^L(x_L) = c_L$$

for any $x_L \geq_{K_2} c_L$ with $x_i > 0$ for $i \in P$. Therefore, if $x \in \Omega_3$, then

$$\lim_{t\to+\infty} \phi_t(x) = \lim_{t\to+\infty} (\phi_t^L(x_L), 0) = (c_L, 0) = c.$$

Define another set

$$\begin{aligned}\Omega_4 = \{&x \in R_+^n : 0 < x_i \le c_i \quad \text{for} \quad i = 1, \dots, r_1, \quad x_i = 0 \quad \text{for} \quad i = r_1 + 1, \dots, k,\\ &x_j \ge c_j \quad \text{for} \quad j = k+1, \dots, n\}\\ &= (c - K) \cap R_+^n \cap \{x \in R_+^n : x_P > 0\}.\end{aligned}$$

We also consider the subsystems $(\mathrm{S_L})$ and $(\mathrm{S_L^*})$ where $L = \{1, \dots, r_1, k+1, \dots, n\}$. By using the same process as in Case (i), we have that

$$\operatorname{diag}(x_L) f_L(x_L, 0) \ge_{K_3} \operatorname{diag}(x_L)(r_L + M_L x_L) \quad \text{for} \quad x_L \in R_+^{n-k+r_1} \quad \text{with} \quad x_L \le_{K_3} c_L,$$

and

$$c_L \ge_{K_3} \phi_t^L(x_L) \ge_{K_3} \psi_t^L(x_L) \qquad \text{for} \quad x_L \le_{K_3} c_L \quad \text{and} \quad t \ge 0,$$

where $r_L = f_L(c_L, 0) - M_L c_L$ and $K_3 = R_+^{r_1} \times (-R_+^{n-k})$.

In a similar way, one easily deduces that

$$\lim_{t \to +\infty} \phi_t^L(x_L) = \lim_{t \to +\infty} \psi_t^L(x_L) = c_L$$

for any $x_L \le_{K_3} c_L$ with $x_i > 0$ for $i \in P$.

If $x \in \Omega_4$ which implies that $x_L \le_{K_3} c_L$ with $x_i > 0$ for $i \in P$, then

$$\lim_{t \to +\infty} \phi_t(x) = \lim_{t \to +\infty} (\phi_t^L(x_L), 0) = (c_L, 0) = c.$$

Therefore, c attracts all points in Ω_3 and Ω_4 with $x_P > 0$. For any $x \in R_+^n$ with $x_P > 0$, just as in Case (i), we can choose $y \in \Omega_3$ and $z \in \Omega_4$ such that

$$z \le_K x \le_K y$$

by using the same method as in (3.6) and (3.7) if $\bar{x}$ is replaced by c. Then

$$\phi_t(z) \le_K \phi_t(x) \le_K \phi_t(y) \quad \text{for} \quad t \ge 0.$$

Since

$$\lim_{t \to +\infty} \phi_t(z) = \lim_{t \to +\infty} \phi_t(y) = c,$$

we conclude that

$$\lim_{t \to +\infty} \phi_t(x) = c$$

for all x in $\{x \in R_+^n : x_P > 0\}$. The proof of sufficiency is completed.

Necessity. Assume that $c \in \partial R_+^n$ attracts all points in $\{x \in R_+^n : x_i > 0,\ i \in P\}$ with $P = \{i : c_i > 0\}$. Then a linearized stability analysis of the steady state c begins with the coefficient matrix of the variational equation given below

$$DF(c) = \operatorname{diag}(f(c)) + \operatorname{diag}(c) Df(c). \tag{4.4}$$

It is easy to see that the necessary condition for c to be stable is $s(DF(c)) \le 0$, which implies that $f(c) \le 0$ by further analyzing equation (4.4). The proof of necessity is completed. □

THEOREM 4.3. *Suppose that the assumption* (KM) *and the condition* (C_1) *are satisfied. Then there exists a steady state* $c \in R^n_+$ *with* $c_P > 0$, $c_{\bar{P}} = 0$ *such that it attracts all solutions with initial conditions in* $\{x \in R^n_+ : x_P > 0\}$.

Before proceeding to the proof of Theorem 4.3, we present a proposition and some lemmas.

PROPOSITION 4.4. *Suppose the assumptions* (KM) *and* (C_1) *hold. Then every solution of* (S) *is bounded.*

Proof. In order to prove that $\phi_t(x)$ is bounded, we only have to prove that $\phi^I_t(x_I)$ and $\phi^J_t(x_J)$ are bounded by Proposition 4.2. Since $\phi^I_t(x_I)$ and $\phi^J_t(x_J)$ are solutions of cooperative systems (S_I) and (S_J), their boundedness can be proved as the proof of Proposition 3.2 in [12] where $Df(0)$ is replaced by M_I and M_J, respectively. □

LEMMA 4.5. *Suppose that* (S) *has no positive steady state. If it has a steady state* $c \in \partial R^n_+$ *with* $\#c = n - 1$, *then Theorem* 4.3 *holds.*

Proof. We may assume $c = (0, c_2, \ldots, c_n)$. If $f_1(c) \le 0$, then the conclusion of Theorem 4.3 is true by Theorem 4.1.

Suppose the contrary; in other words, $f_1(0, c_2, \ldots, c_n) > 0$. Then there exists a sufficiently small $x_1 > 0$ such that $f_1(x_1, c_2, \ldots, c_n) > 0$ by the continuity of f_1. If $i \in I/\{1\}$, then $f_i(x_1, c_2, \ldots, c_n) \ge 0$ because $\frac{\partial f_i}{\partial x_1} \ge 0$ for $i \in I/\{1\}$. If $j \in J$, then $f_j(x_1, c_2, \ldots, c_n) \le 0$ because $\frac{\partial f_j}{\partial x_1} \le 0$ for $j \in J$. Thus, $f(x_1, c_2, \ldots, c_n) \ge_K 0$.

Let $p = (x_1, c_2, \ldots, c_n)$ and $\phi_t(p)$ be a solution of (S) passing through p. Theorem 2.2 may be applied to conclude that $\{\phi_t(p)\}_i$ is nondecreasing for $i \in I$, $t \ge 0$ and $\{\phi_t(p)\}_j$ is nonincreasing for $j \in J$ and $t \ge 0$. Proposition 4.4 shows that $\phi_t(p)$ is bounded. Then we deduce that $\lim_{t\to+\infty} \phi_t(p) = q$. Clearly, q is a steady state of (S). From the type K monotonicity of $\phi_t(p)$ and differential equations (S), we obtain that

$$f_i(\phi_t(p)) \ge 0 \quad \text{for} \quad i \in I, \quad t \ge 0,$$

and

$$f_j(\phi_t(p)) \le 0 \quad \text{for} \quad j \in J, \quad t \ge 0.$$

As $t \to +\infty$, it follows that

$$f_i(q) \ge 0 \quad \text{for} \quad i \in I,$$

and

$$f_j(q) \le 0 \quad \text{for} \quad j \in J.$$

It is easy to see that $q_1 > x_1$, $q_i \ge c_i > 0$ for $i = 2, 3, \ldots, k$. Hence

$$f_i(q) = 0 \quad \text{for} \quad i \in I,$$

which implies that $f(q) \le 0$. Applying Theorem 4.1, we conclude that q attracts all points in $\{x \in R^n_+ : x_Q > 0\}$, where $Q = \{i : q_i > 0\}$. Therefore, Theorem 4.3 is true. □

LEMMA 4.6. *If* $n = 2$, *then the conclusion of Theorem* 4.3 *holds.*

Proof. If (S) has a positive steady state or $c = 0$ is a unique steady state, then by Theorem 3.1 and Theorem C in [13], Theorem 4.3 is true. If there is a steady state which is different from $c = 0$ on the "x_1-axis" or "x_2-axis," then Lemma 4.5 shows that Theorem 4.3 holds. □

Proof of Theorem 4.3. If a positive steady state exists, then the conclusion follows from Theorem 3.1. If $c = 0$ is a unique steady state of (S), then 0_I, 0_J are unique steady states of subsystems (S_I) and (S_J), respectively. By Proposition 4.4, $\phi_t^I(x_I)$ and $\phi_t^J(x_J)$ are bounded. Hence $\lim_{t\to+\infty}\phi_t^I(x_I) = 0_I$ for $x_I \in R_+^k$ and $\lim_{t\to+\infty}\phi_t^J(x_J) = 0_J$ for $x_J \in R_+^{n-k}$ by Theorem C in [13]. Then we obtain that $\lim_{t\to+\infty}\phi_t(x) = 0$ for $x \in R_+^n$ by Proposition 4.2.

If $E \cap \mathrm{Int}R_+^n = \phi$ and $E \neq \{0\}$, we shall prove this theorem by induction on the dimension n of type K monotone systems.

The statement is clear if $n = 2$ by Lemma 4.6. So assume for induction that the theorem is true for all type K systems with dimension less than n. Now, we consider an n-dimensional type K monotone system (S) satisfying (KM) and (C_1). Then we only have to consider the case that there is at least a steady state $c \neq 0$. We may assume that $c \in \{(x_1, x_2, \ldots, x_{n-1}, 0) \in R_+^n : x_i \geq 0 \quad \text{for} \quad i = 1, 2, \ldots, n-1\}$. Let $L = \{1, 2, \ldots, n-1\}$, $\#L = n-1$. We consider the subsystem (S_L)

$$\dot{x}_L = \mathrm{diag}(x_L) f_L(x_L, 0), \quad x_L \in R_+^{n-1}.$$

The condition (C_1) implies that $Df_L(x_L, 0) \leq_{K_4} M_L$ with $s(M_L) < 0$, where $K_4 = R_+^k \times (-R_+^{n-k-1})$. The induction assumption applies to the $(n-1)$-dimensional type K system (S_L). Then (S_L) has a steady state $p = (p_1, p_2, \ldots, p_{n-1})$ which attracts all points in $\{x_L \in R_+^{n-1} : x_Q > 0\}$ where $Q = \{i \in L, p_i > 0\}$. Then it follows that $\lim_{t\to+\infty}\phi_t^L(x_L) = p$ for $x_L \in \{x \in R_+^{n-1} : x_Q > 0\}$.

By the necessity of Theorem 4.1, we have

$$f_i(p_1, \ldots, p_{n-1}, 0) \leq 0 \quad \text{for} \quad i = 1, 2, \ldots, n-1.$$

If $f_n(p_1, \ldots, p_{n-1}, 0) \leq 0$, then $f(p_1, \ldots, p_{n-1}, 0) \leq 0$ and Theorem 4.3 is true by the sufficiency of Theorem 4.1. Otherwise, $f_n(p_1, \ldots, p_{n-1}, 0) > 0$ holds. Define $J_1 = \{j \in J : p_j > 0\} \cup \{n\}$. Then

$$f_j(p_1, \ldots, p_{n-1}, 0) \leq 0 \quad \text{for} \quad j \in J/J_1,$$

and

$$f_j(p_1, \ldots, p_{n-1}, 0) = 0 \quad \text{for} \quad j \in J_1/\{n\}.$$

Let $L_1 = I \cup J_1$ and consider the subsystem (S_{L_1})

$$\dot{x}_{L_1} = \mathrm{diag}(x_{L_1}) f_{L_1}(x_{L_1}, 0), \quad x_{L_1} \in R_+^{\#L_1},$$

where $x_{L_1} = (x_I, x_{J_1})$, $0 = 0_{J/J_1}$. By the continuity of f_n, there exists a sufficiently small $x_n > 0$ such that $f_n(p_1, \ldots, p_{n-1}, x_n) > 0$. If $j \in J_1/\{n\}$, then $f_j(p_1, \ldots, p_{n-1}, 0) = 0$ and $\frac{\partial f_j}{\partial x_n} \geq 0$, which implies that $f_j(p_1, \ldots, p_{n-1}, x_n) \geq 0$. If $i \in I$, $f_i(p_1, \ldots, p_{n-1}, 0) \leq 0$ and $\frac{\partial f_i}{\partial x_n} \leq 0$ which implies that $f_i(p_1, \ldots, p_{n-1}, x_n) \leq 0$. Let $v = (p_1, \ldots, p_{n-1}, x_n)$. Then $v = (p_I, p_{J_1/\{n\}}, 0_{J/J_1}, x_n) = (u, 0)$, where $u = (p_I, p_{J_1/\{n\}}, x_n) \in R_+^{\#L_1}$. Thus $f_{L_1}(u, 0) \leq_{K_5} 0$ where $K_5 = R_+^k \times (-R_+^{\#J_1})$. Theorem 2.2 may be applied to conclude that $\{\phi_t^{L_1}(u)\}_i$ is nonincreasing if $i \in I$ and $\{\phi_t^{L_1}(u)\}_j$ is nondecreasing if $j \in J_1$ for $t \geq 0$. Also, $\phi_t^{L_1}(u)$ is bounded by Proposition 4.4. Thus we obtain that $\lim_{t\to+\infty}\phi_t^{L_1}(u) = q_{L_1}$. Clearly, q_{L_1} is a steady state of (S_{L_1}).

On the other hand, the type K monotonicity of $\phi_t^{L_1}(u)$ implies that for any $t \geq 0$,

$$f_i(\phi_t^{L_1}(u), 0) \leq 0 \qquad \text{if} \quad i \in I,$$

and

$$f_j(\phi_t^{L_1}(u), 0) \geq 0 \qquad \text{if} \quad j \in J_1.$$

As $t \to +\infty$, we have

$$f_i(q_{L_1}, 0) \leq 0 \quad \text{for} \quad i \in I,$$

and

$$f_j(q_{L_1}, 0) \geq 0 \quad \text{for} \quad j \in J_1.$$

Since $q_n > x_n$, $q_j \geq p_i > 0$ for $j \in J_1/\{n\}$, it follows that

$$f_j(q_{L_1}, 0) = 0 \quad \text{for} \quad j \in J_1.$$

If $f_j(q_{L_1}, 0) \leq 0$ for $j \in J/J_1$, then the theorem holds by Theorem 4.1. Assume it is false. We define $J_2 = \{j \in J : f_j(q_{L_1}, 0) > 0\}$. Let $L_2 = I \cup J_1 \cup J_2$ and consider the subsystem $(\mathrm{S}_{\mathrm{L}_2})$

$$\dot{x}_{L_2} = \mathrm{diag}(x_{L_2}) f_{L_2}(x_{L_2}, 0), \quad x_{L_2} \in R_+^{\#L_2},$$

where $x_{L_2} = (x_I, x_{J_1}, x_{J_2}), 0 = 0_{J/(J_1 \cup J_2)}$.

Since

$$f_{J_2}(q_{L_1}, 0_{J_2}, 0_{J/(J_1 \cup J_2)}) > 0,$$

$$f_{J_1}(q_{L_1}, 0_{J_2}, 0_{J/(J_1 \cup J_2)}) = 0,$$

and

$$f_I(q_{L_1}, 0_{J_2}, 0_{J/(J_1 \cup J_2)}) \leq 0$$

hold, there exists $x_{J_2} > 0$ small enough such that $f_{J_2}(q_{L_1}, x_{J_2}, 0_{J/(J_1 \cup J_2)}) > 0$ by the continuity of f_{J_2}. Meanwhile, $\frac{\partial f_{J_1}}{\partial x_{J_2}} \geq 0$ implies that $f_{J_1}(q_{L_1}, x_{J_2}, 0_{J/(J_1 \cup J_2)}) \geq 0$, and $\frac{\partial f_I}{\partial x_{J_2}} \leq 0$ implies that $f_I(q_{L_1}, x_{J_2}, 0_{J/(J_1 \cup J_2)}) \leq 0$ for such x_{J_2}. Hence, $f_{L_2}(q_{L_1}, x_{J_2}, 0_{J/(J_1 \cup J_2)}) \leq_{K_6} 0$, where $K_6 = R_+^k \times (-R_+^{\#(J_1 \cup J_2)})$. Let $v = (q_{L_1}, x_{J_2})$. Then $\{\phi_t^{L_2}(v)\}_i$ is nonincreasing for $i \in I$ and $\{\phi_t^{L_2}(v)\}_j$ is nondecreasing for $j \in J_1 \cup J_2$ on $[0, +\infty)$ by Theorem 2.2. Because $\phi_t^{L_2}(v)$ is bounded, it deduces that $\lim_{t \to +\infty} \phi_t^{L_2}(v) = r_{L_2}$. Obviously, r_{L_2} is a steady state of $(\mathrm{S}_{\mathrm{L}_2})$. The type K monotonicity of $\phi_t^{L_2}(v)$ shows that for $t \geq 0$,

$$f_i(\phi_t^{L_2}(v), 0_{J/J_1 \cup J_2}) \leq 0 \qquad \text{if} \quad i \in I,$$

and

$$f_j(\phi_t^{L_2}(v), 0_{J/J_1 \cup J_2}) \geq 0 \qquad \text{if} \quad j \in J_1 \cup J_2.$$

Hence, as $t \to +\infty$, we have

$$f_i(r_{L_2}, 0) \le 0 \qquad \text{for} \quad i \in I,$$

and

$$f_j(r_{L_2}, 0) \ge 0 \qquad \text{for} \quad j \in J_1 \cup J_2.$$

Evidently, $r_{J_2} > x_{J_2}$, $r_{J_1} \ge q_{J_1} > 0$. Then

$$f_j(r_{L_2}, 0) = 0 \quad \text{for} \quad j \in J_1 \cup J_2.$$

If $j \in J/(J_1 \cup J_2)$, $f_j(r_{L_2}, 0) \le 0$ which is equivalent to $f(r_{L_2}, 0) \le 0$, then the theorem is true. If not, we define $J_3 = \{j \in J : f_j(r_{L_2}, 0) > 0\}$. Let $L_3 = I \cup J_1 \cup J_2 \cup J_3$ and discuss the subsystem $(\mathrm{S}_{\mathrm{L}_3})$ in a similar way. This process must be stopped at most at the $(n-k)$th step.

Hence, in this way, we can find a steady state c satisfying $f(c) \le 0$. Then by Theorem 4.1, c attracts all points in $\{x \in R^n_+ : x_Q > 0\}$ where $Q = \{i \in N, c_i > 0\}$. Therefore we have proved that the conclusion of Theorem 4.3 holds for n-dimensional type K monotone systems. This completes the proof. □

In order to give the first question in the introduction another answer, we present a different condition. Suppose that $(\mathrm{S_I})$ and $(\mathrm{S_J})$ have positive steady states x^0_1 and x^0_2. Let $p = (x^0_1, 0)$ and $q = (0, x^0_2)$. The system (S) is called to satisfy the condition (C'_2) if either

$$Df(x) \le_K Df(y) \quad \text{whenever} \quad q \le_K y \le_K x \le_K p \tag{4.5}$$

or

$$Df(x) \le_K Df(y) \quad \text{whenever} \quad q \le_K x \le_K y \le_K p. \tag{4.6}$$

THEOREM 4.7. *Let the hypotheses* (KM) *and* (C'_2) *hold. If* $(\mathrm{S_I})$ *and* $(\mathrm{S_J})$ *have steady states* x^0_1 *and* x^0_2 *which are globally asymptotically stable in* $\mathrm{Int}R^k_+$ *and* $\mathrm{Int}R^{n-k}_+$, *respectively, and each is stable to invasion by every competing species*

$$f_J(x^0_1, 0) \le 0$$

and

$$f_I(0, x^0_2) \le 0,$$

then (S) *cannot have a stable positive steady state.*

Proof. We only have to prove the case that the system (S) satisfies (4.5). Another case can be proved in the same method. In fact, suppose it is false. Then there exists at least a stable positive steady state $\bar{x}$ and $\bar{x} \le_K (x^0_1, 0) = p$ by Proposition 3.3 in [5, p. 864]. The calculation shows that

$$f(\bar{x}) - f(p) = \left[\int_0^1 Df(s\bar{x} + (1-s)p)ds\right](\bar{x} - p).$$

We denote the matrix in brackets by U. The condition (4.5) implies that $U \le_K Df(\bar{x})$. Because $\bar{x}$ is stable, $s(Df(\bar{x})) < 0$. It follows that $s(U) < 0$ from (i) of Theorem 2.3 and $(-U)^{-1} \ge_K 0$ from (iii) of Theorem 2.3. So $(-U)^{-1} f(p) \ge_K 0$. But $\bar{x} \le_K p$

implies that $(-U)^{-1}f(p) = (\bar{x} - p) \leq_K 0$ which implies that $\bar{x} - p = 0$ from the just-proved fact. However, $\bar{x} \neq p$. This is a contradiction. Therefore, (S) cannot have a stable positive steady state. □

Before finishing this section, we remark that Theorems 4.1 and 4.3 together with Theorem 3.1 completely generalize the results of Takeuchi and Adachi [6] to more general Kolmogorov systems with the conditions (KM) and (C_1). Theorem 4.1 implies that every saturated steady state is globally asymptotically stable and that it is not possible for both systems (S_I) and (S_J) to have positive steady states each of which cannot be invaded by the competing subcommunity as long as (KM) and (C_1) are satisfied. Therefore, these results provide an answer to each of the questions posed by Smith [5] which have been mentioned in the introduction. Theorem 4.7 gives the first question another answer.

5. The permanence for system (S). Consider the system

$$\dot{x}_i = x_i f_i(x_1, x_2, \ldots, x_n), \quad 1 \leq i \leq n, \quad x_i \geq 0, \tag{S}$$

which satisfies the assumption (KM).

In the paper [5], Smith exhibited sufficient conditions for the permanence of (S). His essential condition is

$$\text{(H)} \quad \begin{cases} (S_I) \text{ possesses a positive steady state } x_1^0 \text{ which is unstable to } R_+^{n-k} \\ \quad \text{(that is, } f_J(x_1^0, 0) > 0), \\ (S_J) \text{ possesses a positive steady state } x_2^0 \text{ which is unstable to } R_+^k \\ \quad \text{(that is, } f_I(0, x_2^0) > 0). \end{cases}$$

His additional assumption is

$$\text{(AH)} \quad \begin{cases} x_1^0 + R_+^k \text{ lies in the domain of attraction of } x_1^0 \text{ for } (S_I), \\ x_2^0 + R_+^{n-k} \text{ lies in the domain of attraction of } x_2^0 \text{ for } (S_J). \end{cases}$$

Under the hypothesis (H), Smith concluded that there exist positive steady states x^1 and x^2 of (S) satisfying that $0 < x^1, x^2 \leq (x_1^0, x_2^0)$, and $x^1 \leq_K x^2$ (see [5, Theorem 3.6]). Meanwhile, combining the hypothesis (H) with (AH), he deduced that $\omega(x) \subset [x^1, x^2]_K$ for all $x > 0$.

Recall that the system (S) is permanent if there exist constant numbers $\delta, D > 0$ such that for every solution $x(t)$ of (S) with $x(0) > 0$, $\delta \leq \liminf_{t \to +\infty} x_i(t) \leq \limsup_{t \to +\infty} x_i(t) \leq D$ for all $i \in N$. Then the system (S) is permanent under the hypotheses (H) and (AH).

Clearly, the species satisfying (H) is nonobligate (that is, $f(0) > 0$). In obligate cases, the authors gave the sufficient conditions to guarantee that the system (S) is permanent (see [10, Theorem 2.1]). In those situations, two subsystems (S_L) and (S_P) play an important role, where $L \subset N, L \supset I$ and $P \subset N, P \supset J$. Our essential condition is

$$\text{(H}_1\text{)} \quad \begin{cases} x_L^0 \text{ is a positive steady state of } (S_L),\ f_{\bar{L}}(x_L^0, 0) > 0, \\ x_P^0 \text{ is a positive steady state of } (S_P),\ f_{\bar{P}}(0, x_P^0) > 0. \end{cases}$$

Our additional assumption is

$$\text{(AH}_1\text{)} \quad \begin{cases} x_L^0 + K_1 \text{ lies in the domain of attraction of } x_L^0 \text{ for } (S_L), \\ x_P^0 - K_2 \text{ lies in the domain of attraction of } x_P^0 \text{ for } (S_P), \end{cases}$$

where $K_1 = R_+^k \times (-R_+^{l-k})$, $K_2 = R_+^{p+k-n} \times (-R_+^{n-k})$. We have the following result.

THEOREM 5.1. *Let the assumption* (H_1) *hold and* $(0, x_P^0) \leq_K (x_L^0, 0)$. *Then there exist positive steady states* $\bar{x}$ *and* $\widetilde{x}$ *for* (S) *satisfying* $0 < \bar{x}, \widetilde{x} \leq ((x_L^0)_I, (x_P^0)_J)$ *and* $\bar{x} \leq_K \widetilde{x}$. *In addition, if the assumption* (AH_1) *holds, then* $\omega(x) \subset [\bar{x}, \widetilde{x}]_K$ *for all* $x > 0$. *And* $\omega(x) = \{\bar{x}\}$ *for all* $x > 0$ *with* $x \leq_K \bar{x}$ *and* $\omega(x) = \{\widetilde{x}\}$ *for all* $x > 0$ *with* $x \geq_K \widetilde{x}$.

In Theorem 5.1, $(x_L^0)_I$, $(x_P^0)_J$ are components $(x_L^0)_i$ of x_L^0 for all $i \in I$ and $(x_P^0)_j$ of x_P^0 for all $j \in J$, respectively. These results generalize Smith's Theorem 3.6 in [5].

In the same paper, Smith also obtained the following result (see [5, Theorem 3.8]) under the assumption

$$(AH_2) \qquad \begin{cases} \frac{\partial f_I}{\partial x_I}(x_I, 0) \geq \frac{\partial f_I}{\partial x_I}(\bar{x}_I, 0), & 0 \leq x_I \leq \bar{x}_I, \\ \frac{\partial f_J}{\partial x_J}(0, x_J) \geq \frac{\partial f_J}{\partial x_J}(0, \bar{x}_J), & 0 \leq x_J \leq \bar{x}_J. \end{cases}$$

THEOREM 5.2. *Let the assumptions* (H) *and* (AH_2) *hold. Then* x_1^0 *is globally asymptotically stable for* (S_I) *with respect to* $\mathrm{Int}R_+^k$ *and* x_2^0 *is globally asymptotically stable for* (S_J) *with respect to* $\mathrm{Int}R_+^{n-k}$ *and there are positive steady states* $x^1 \leq_K x^2$. *If* $x > 0$, *then* $\omega(x) \subset [x^1, x^2]_K$. x^1 *attracts points* $x > 0$ *with* $x \leq_K x^1$ *and* x^2 *attracts points* $x > 0$ *with* $x \geq_K x^2$.

Then the system (S) is permanent in the case when (H) and (AH_2) hold. However, as Smith pointed out, "there may be several stable steady states representing persistence. It is easy to see that one cannot prove uniqueness of a positive stable steady state under the hypotheses of our Theorems 3.6 and 3.8 by simply considering two-dimensional competitive systems." "An interesting open problem is to give sufficient conditions for the uniqueness of a positive steady state in the context of Theorem 3.6 or 3.8" (see [5, p. 869]). One of the authors in [14] provided a sufficient condition such that a positive steady state is unique in the context of Smith's Theorem 3.6. His additional conditions are

$$s\left(\frac{\partial f_J}{\partial x_J}(0, x_2^0)\right) < 0$$

and

$$Df(y) \leq_K Df(x) \quad \text{whenever} \quad x, y \in [0, (x_1^0, x_2^0)] \quad \text{with} \quad x \leq_K y.$$

Together with Theorems 5.2 and 3.1, we know that (S) has a globally asymptotically stable steady state if the assumptions (H), (AH_2), and (C_1) hold. In this paper, as another application of Theorem 3.1, we give a sufficient condition for the uniqueness of a positive steady state.

THEOREM 5.3. *Let the hypotheses* (C_1) *and* (H_1) *hold. Then* (S) *has a globally asymptotically stable positive steady state.*

Proof. We only have to show that Theorem 5.3 satisfies that $(0, x_P^0) \leq_K (x_L^0, 0)$. Since x_L^0 is a positive steady state of (S_L) and x_P^0 is a positive steady state of (S_P), it follows that

$$\lim_{t \to +\infty} \phi_t^L(x_L) = x_L^0 \quad \text{for} \quad x_L \in \mathrm{Int}R_+^{\#L}$$

and

$$\lim_{t \to +\infty} \phi_t^P(x_P) = x_P^0 \quad \text{for} \quad x_P \in \mathrm{Int}R_+^{\#P}$$

from Theorem 3.1. It is easy to choose $x_L \in \mathrm{Int}R_+^{\#L}$ and $x_P \in \mathrm{Int}R_+^{\#P}$ such that $(0, x_P) \leq_K (x_L, 0)$. Theorem 2.1 shows that

$$\phi_t(0, x_P) \leq_K \phi_t(x_L, 0) \qquad \text{for} \quad t \geq 0;$$

that is,

$$(0, \phi_t^P(x_P)) \leq_K (\phi_t^L(x_L), 0) \qquad \text{for} \quad t \geq 0.$$

Then

$$(0, x_P^0) \leq_K (x_L^0, 0).$$

Applying Theorem 5.1, we conclude that (S) has a positive steady state and Theorem 3.1 implies that the positive steady state is globally asymptotically stable with respect to $\mathrm{Int}R_+^n$. □

Finally, we shall present the permanence result for system (S). Let us introduce the assumption

$$(\mathrm{AH}_3) \qquad \begin{cases} Df_L(x_L, 0) \leq_{K_1} M_L & \text{and} \quad s(M_L) < 0 \quad \text{for} \quad x_L \in R_+^l, \\ Df_P(0, x_P) \leq_{K_2} M_P & \text{and} \quad s(M_P) < 0 \quad \text{for} \quad x_P \in R_+^p, \end{cases}$$

where $N \supset L \supset I$, $N \supset P \supset J$, $l = \#L$, $p = \#P$, $K_1 = R_+^k \times (-R_+^{l-k})$, and $K_2 = R_+^{p+k-n} \times (-R_+^{n-k})$. Under the hypothesis (AH_3), if $(\mathrm{S_L})$ and $(\mathrm{S_P})$ have positive steady states x_L^0 and x_P^0, respectively, then x_L^0 is globally stable for $(\mathrm{S_L})$ with respect to $\mathrm{Int}R_+^l$ and x_P^0 is globally stable for $(\mathrm{S_P})$ with respect to $\mathrm{Int}R_+^p$ by Theorem 3.1. Hence, combining Theorems 5.1 and 3.1, we obtain the following permanence result.

THEOREM 5.4. *Let the hypotheses* (H_1) *and* (AH_3) *hold. Then* x_L^0 *is globally stable for* $(\mathrm{S_L})$ *with respect to* $\mathrm{Int}R_+^l$ *and* x_P^0 *is globally stable for* $(\mathrm{S_P})$ *with respect to* $\mathrm{Int}R_+^p$*, and there are positive steady states* $\bar{x} \leq_K \widetilde{x}$ *such that* $\omega(x) \subset [\bar{x}, \widetilde{x}]_K$ *for* $x > 0$. $\bar{x}$ *attracts points* $x > 0$ *with* $x \leq_K \bar{x}$ *and* $\widetilde{x}$ *attracts points* $x > 0$ *with* $x \geq_K \widetilde{x}$.

REFERENCES

[1] M. W. HIRSCH, *Systems of differential equations which are competitive or cooperative* I: *Limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.

[2] M. W. HIRSCH, *Systems of differential equations that are competitive or cooperative* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 423–439.

[3] M. W. HIRSCH, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–53.

[4] H. L. SMITH, *Monotone Dynamical Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.

[5] H. L. SMITH, *Competing subcommunities of mutualists and a generalized Kamke theorem*, SIAM J. Appl. Math., 46 (1986) pp. 856–874.

[6] Y. TAKEUCHI AND N. ADACHI, *The existence of globally stable equilibria of ecosystems of the generalized Volterra type*, J. Math. Biol., 10 (1980), pp. 401–415.

[7] Y. TAKEUCHI, N. ADACHI, AND H. TOKUMARU, *Global stability of ecosystems of the generalized Volterra type*, Math. Biosci., 42 (1978), pp. 119–136.

[8] C. C. TRAVIS AND W. M. POST, *Dynamics and comparative statistics of mutualistic communities*, J. Theor. Biol., 78 (1979), pp. 553–571.

[9] B. S. GOH, *Stability in models of mutualism*, Amer. Naturalist, 113 (1979), pp. 261–275.

[10] C. F. TU AND J. F. JIANG, *The coexistence of a community of species with limited competition*, J. Math. Anal. Appl., 217 (1998), pp. 233–245.

[11] J. F. SELGRADE, *Asymptotic behavior of solutions to single loop positive feedback systems*, J. Differential Equations, 38 (1980), pp. 80–103.

[12] J. F. Jiang, *Convergence in cooperative models with a weak concavity*, Chinese Ann. Math. Ser. A (1999).
[13] J. F. Jiang, *On the global stability of cooperative systems*, Bull. London Math. Soc., 26 (1994), pp. 455–458.
[14] J. F. Jiang, *The global stability of a system modeling a community with limited competition*, Proc. Amer. Math. Soc., 125 (1997), pp. 1381–1389.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 379–400

ANALYTIC REGULARITY FOR A SINGULARLY PERTURBED PROBLEM*

J. M. MELENK† AND C. SCHWAB†

Abstract. A singularly perturbed equation of elliptic-elliptic type in two dimensions is considered. We assume analyticity of the input data, i.e., the boundary of the domain is an analytic curve, the boundary data are analytic, and the right-hand side is analytic. We give asymptotic expansions of the solution and new error bounds that are uniform in the perturbation parameter as well as in the expansion order. Additionally, we provide growth estimates for higher derivatives of the solution where the dependence on the perturbation parameter appears explicitly. These error bounds and growth estimates are used in [J. M. Melenk and C. Schwab, *SIAM J. Numer. Anal.,* 35 (1998), pp. 1520–1557] to construct *hp* versions of the finite element method which feature *robust exponential convergence*, i.e., the rate of convergence is exponential and independent of the perturbation parameter ε.

Key words. boundary layer, singularly perturbed problem, asymptotic expansions, error bounds, Gevrey regularity

AMS subject classifications. 35B25, 35C20, 35B65, 65N30

PII. S0036141097317542

1. Introduction.

Numerous partial differential equation models contain large or small parameters. Of interest in solid mechanics are, for example, the plate and shell equations at small thickness and nearly incompressible solids. The presence of small parameters often implies that the problem is singularly perturbed, and much attention has been devoted in the past decades to the asymptotic analysis of the solution; we mention here only [2, 3]. Typically, the solutions admit decompositions into a smooth part and so-called boundary layers. While the asymptotic structure of the solution is usually known (see, e.g., [4, 5, 6, 7]), the asymptotic expansions are often too complex for practical computations, and one has to resort to finite element solutions of the boundary value problem (BVP) of interest. Here the singular perturbation character of the problem and the boundary layer components of the solution cause approximability problems which often manifest themselves as loss of robustness, i.e., the performance of the numerical method depends strongly on the perturbation parameter. Given stability of a numerical method for a BVP, the key to its convergence is the regularity of the solution, particularly bounds on higher derivatives. In spectral and *hp* finite element methods (*hp* FEMs) that aim at exponential rates of convergence, analytic regularity results, i.e., results concerning the growth of the derivatives of the solution, are crucial. It is the purpose of the present paper to provide such analytic regularity results for the following, singularly perturbed elliptic-elliptic model problem:

$$\begin{aligned} L_\varepsilon u_\varepsilon \equiv -\varepsilon^2 \Delta u_\varepsilon + u_\varepsilon &= f \qquad \text{in } \Omega \subset \mathbb{R}^2, \\ u_\varepsilon &= g \qquad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where $\partial\Omega$ is a closed, nonselfintersecting, analytic curve, f is analytic on a neighborhood of $\overline{\Omega}$, g is analytic on $\partial\Omega$, and $\varepsilon \in (0, 1]$ is a small parameter.

*Received by the editors February 28, 1997; accepted for publication (in revised form) April 29, 1998; published electronically January 5, 1999.
http://www.siam.org/journals/sima/30-2/31754.html
†Seminar für Angewandte Mathematik, ETH Zürich, CH–8092 Zürich, Switzerland (melenk@sam.math.ethz.ch, schwab@sam.math.ethz.ch).

As usual, we denote by $L^2(\Omega)$ the square integrable functions on Ω and by $H^1(\Omega)$ those functions of $L^2(\Omega)$ whose (distributional) derivative is also in $L^2(\Omega)$. The trace operator maps $H^1(\Omega)$ onto the space $H^{1/2}(\partial\Omega)$ by restricting the elements of $H^1(\Omega)$ to the boundary $\partial\Omega$. $H^1_0(\Omega)$ denotes the kernel of the trace operator; that is, it is given by those functions in $H^1(\Omega)$ whose trace on $\partial\Omega$ is zero.

The weak formulation of (1.1) is to find $u_\varepsilon \in H^1(\Omega)$ such that $u_\varepsilon|_{\partial\Omega} = g$ and

$$B_\varepsilon(u_\varepsilon, v) := \varepsilon^2 \int_\Omega \nabla u_\varepsilon \cdot \nabla v \, dxdy + \int_\Omega u_\varepsilon v \, dxdy = F(v) := \int_\Omega fv \, dxdy \tag{1.2}$$

holds for all $v \in H^1_0(\Omega)$. Associated with this problem is the notion of an energy

$$\|u\|^2_{\varepsilon,\Omega} := B_\varepsilon(u,u) = \varepsilon^2 \|\nabla u\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)}$$

and an energy norm, being the square root of the energy. We have the a priori estimate

$$\|u_\varepsilon\|_{\varepsilon,\Omega} \le \|f\|_{L^2(\Omega)} + C\|g\|_{H^{1/2}(\partial\Omega)} \tag{1.3}$$

for some $C > 0$ independent of ε.

As the input data f, g, and $\partial\Omega$ of (1.1) are analytic, standard elliptic regularity theory [8, 9] implies that the exact solution u_ε is analytic on a neighborhood of $\overline{\Omega}$, i.e., it satisfies estimates of the form

$$\|D^\alpha u_\varepsilon\|_{L^\infty(\Omega)} \le |\alpha|! C_\varepsilon K_\varepsilon^{|\alpha|} \qquad \forall \alpha = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2,$$

where $\mathbb{N}$, $\mathbb{N}_0$ denote the set of positive and nonnegative integers, respectively. However, the constants C_ε and K_ε depend on ε in an unspecified way and our aim here is to control explicitly the dependence of the derivatives of u_ε on the perturbation parameter ε. By carefully tracking the ε-dependence of the constants in the standard results, we obtain in section 3 (Theorem 3.1) the following estimate:

$$\|D^\alpha u_\varepsilon\|_{L^2(\Omega)} \le C K^{|\alpha|} \max(|\alpha|, \varepsilon^{-1})^{|\alpha|} \qquad \forall \alpha \in \mathbb{N}_0^2 \tag{1.4}$$

with C, $K > 0$ independent of ε. Note that for $|\alpha| \ge \varepsilon^{-1}$ this yields an estimate independent of ε. This estimate is also sufficient to prove that polynomials of degree p can approximate the solution u_ε at a robust exponential rate provided that the polynomial degree p is at least $O(\varepsilon^{-1})$.

For derivatives of order $|\alpha| < \varepsilon^{-1}$, the estimates (1.4) are too pessimistic in that they do not capture accurately the boundary layer behavior of the solution. Typically, the solutions of (1.1) exhibit boundary layers for small ε; that is, the behavior of the solution normal to the boundary differs substantially from the behavior in the tangential direction. Classically, this anisotropic behavior of boundary layers can be described with asymptotic expansions. In section 2, we therefore provide analytic regularity results for the terms of the asymptotic expansion and estimates on the remainder (Theorem 2.2). The new feature of our results over the classical assertions for asymptotic expansions is that in all our estimates, the dependence on ε and the expansion order M is made explicit. This precise control over the asymptotic expansions is an essential ingredient in the design and convergence analysis of an *hp* FEM for (1.1) that converges at a robust exponential rate [1].

Remark 1.1. One of the difficulties in the analysis of singularly perturbed problems is the variety of phenomena that can arise already in the linear case: boundary layers, internal layers, corner layers, multiple length scales, etc. We analyze the model problem (1.1) with strong assumptions on the data f, g, $\partial\Omega$ in order to be able to concentrate on one of these phenomena, namely boundary layers.

2. Analysis of the asymptotic expansion. In this section we analyze the classical asymptotic expansions for the solution of (1.1). The asymptotic expansions (defined more precisely in section 2.2) allow us to decompose the solution u_ε as

$$u_\varepsilon = w_M + \chi u_M^{BL} + r_M, \tag{2.1}$$

where $M \in \mathbb{N}_0$ indicates the expansion order, w_M is the truncated outer expansion, u_M^{BL} is the truncated inner expansion, χ is a cutoff function supported by a neighborhood of $\partial\Omega$, and r_M is a remainder. Our results concerning w_M, u_M^{BL}, and r_M are collected in section 2.3, Theorem 2.2: We give analytic regularity results for w_M and u_M^{BL} that are uniform in ε and M, and we show that u_M^{BL} (together with all its derivatives) decays exponentially normal to $\partial\Omega$. Furthermore, we give new error bounds for the remainder r_M which are explicit in ε and the expansion order M.

2.1. Notation. We introduce *boundary fitted coordinates* to define the asymptotic expansions of the exact solution. Let $(X(\theta), Y(\theta))$, $\theta \in [0, L)$, be an analytic, L-periodic parametrization of the boundary $\partial\Omega$ by arc length such that the normal vector $(-Y'(\theta), X'(\theta))$ always points into the domain Ω. Introduce the notation $\kappa(\theta)$ for the curvature of the boundary curve and denote by $\mathbb{T}_L$ the one-dimensional torus of length L, i.e., $\mathbb{R}/L\mathbb{Z}$, endowed with the usual topology. The functions X, Y, and hence also κ are analytic on $\mathbb{T}_L$ by the analyticity of $\partial\Omega$. For the remainder of this section, let $\rho_0 > 0$ be fixed such that

$$0 < \rho_0 < \frac{1}{\|\kappa\|_{L^\infty(\mathbb{T}_L)}}. \tag{2.2}$$

Then the mapping

$$\begin{array}{rcl} \psi : [0,\rho_0] \times \mathbb{T}_L & \to & \overline{\Omega}, \\ (\rho, \theta) & \mapsto & (X(\theta) - \rho Y'(\theta), Y(\theta) + \rho X'(\theta)) \end{array} \tag{2.3}$$

is real analytic on (a neighborhood of) $[0, \rho_0] \times \mathbb{T}_L$. The function ψ maps the rectangle $(0, \rho_0) \times \mathbb{T}_L$ onto a half-tubular neighborhood Ω_0 of $\partial\Omega$. Furthermore, by the choice of ρ_0, the inverse $\psi^{-1} : \overline{\Omega}_0 \to [0, \rho_0] \times \mathbb{T}_L$ exists and is also real analytic on (a neighborhood of) the closed set $\overline{\Omega}_0$. For technical reasons we will be able to define the boundary layer expansion (the inner expansion) only in a neighborhood of the boundary $\partial\Omega$. Therefore, we introduce a cutoff function χ supported by a neighborhood of $\partial\Omega$. For ease of notation, let us define χ in the neighborhood of $\partial\Omega$ in boundary fitted coordinates (ρ, θ). Fix

$$0 < \rho_1 < \rho_0, \tag{2.4}$$

and let χ be a smooth cutoff function, defined on $[0, \infty) \times \mathbb{T}_L$, satisfying

$$\chi = \begin{cases} 1 \text{ for } 0 \le \rho \le \rho_1, \\ 0 \text{ for } \rho \ge (\rho_1 + \rho_0)/2. \end{cases} \tag{2.5}$$

Finally, as the right-hand side f of (1.1) is assumed to be analytic on (a neighborhood of) $\overline{\Omega}$, there is a complex neighborhood $\tilde{\Omega} \subset \mathbb{C} \times \mathbb{C}$ of $\overline{\Omega}$ and a holomorphic extension of f (for convenience again denoted by f) to $\tilde{\Omega}$. From Cauchy's integral theorem for derivatives, we therefore have, after passing to a compact subset of $\tilde{\Omega}$ which is again denoted $\tilde{\Omega}$, the existence of constants C_f, $\gamma \ge 0$ such that

$$\|\Delta^{(i)} f\|_{L^\infty(\tilde{\Omega})} \le C_f (2i)! \gamma^{2i} \qquad \forall i \in \mathbb{N}_0, \tag{2.6}$$

where $\Delta^{(i)}$ denotes the iterated Laplace operator, i.e., $\Delta^{(0)} = \mathrm{Id}$, $\Delta^{(1)} = \Delta$, $\Delta^{(2)} = \Delta\Delta$, etc.

2.2. Inner and outer expansion. For every $M \in \mathbb{N}_0$ the *outer expansion* of order $2M$ is defined by

$$w_M := \sum_{i=0}^{M} \varepsilon^{2i} \Delta^{(i)} f. \tag{2.7}$$

The function $u_\varepsilon - w_M$ then satisfies

$$L_\varepsilon(u_\varepsilon - w_M) = f - L_\varepsilon w_M = \varepsilon^{2M+2} \Delta^{(M+1)} f. \tag{2.8}$$

So, asymptotically as ε tends to zero, the functions w_M satisfy the differential equation in Ω. However, the functions w_M do not satisfy the given boundary conditions g. We therefore introduce a boundary layer correction u^{BL} of w_M, which will lead to the inner expansion. The correction u^{BL} is defined as the solution of

$$\begin{aligned} L_\varepsilon u^{BL} &= 0 && \text{in } \Omega, \\ u^{BL} &= g - \sum_{i=0}^{M} \varepsilon^{2i} \left[\Delta^{(i)} f \right]_{\partial\Omega} && \text{on } \partial\Omega. \end{aligned} \tag{2.9}$$

The *inner expansion* is an asymptotic expansion for this correction function u^{BL}. In order to define this expansion, we need to rewrite the differential operator L_ε in the boundary fitted coordinates (ρ, θ). With the curvature $\kappa(\theta)$ of $\partial\Omega$ and the function

$$\sigma(\rho, \theta) = \frac{1}{1 - \kappa(\theta)\rho}, \tag{2.10}$$

we have (see, for example, [4])

$$\Delta u(\rho, \theta) = \partial_\rho^2 u - \kappa(\theta)\sigma(\rho, \theta)\partial_\rho u + \sigma^2(\rho, \theta)\partial_\theta^2 u + \rho\kappa'(\theta)\sigma^3(\rho, \theta)\partial_\theta u.$$

Introducing now the *stretched variable* notation $\widehat{\rho} = \rho/\varepsilon$, we have

$$L_\varepsilon = -\partial_{\widehat{\rho}}^2 + \mathrm{Id} + \varepsilon\kappa(\theta)\sigma(\varepsilon\widehat{\rho}, \theta)\partial_{\widehat{\rho}} - \varepsilon^2\sigma^2(\varepsilon\widehat{\rho}, \theta)\partial_\theta^2 - \varepsilon^3\widehat{\rho}\kappa'(\theta)\sigma^3(\varepsilon\widehat{\rho}, \theta)\partial_\theta .$$

Expanding in power series in ε, we can write the operator L_ε formally as

$$L_\varepsilon = \sum_{i=0}^{\infty} \varepsilon^i L_i, \tag{2.11}$$

where the operators L_i have the form

$$L_0 = -\partial_{\widehat{\rho}}^2 + \mathrm{Id}, \quad L_i = -\widehat{\rho}^{i-1} a_1^{i-1} \partial_{\widehat{\rho}} - \widehat{\rho}^{i-2} a_2^{i-2} \partial_\theta^2 - \widehat{\rho}^{i-2} a_3^{i-3} \partial_\theta , \quad i \geq 1, \tag{2.12}$$

and the coefficients a_j^i are given by

$$a_1^i = -[\kappa(\theta)]^{i+1}, \; a_2^i = (i+1)[\kappa(\theta)]^i, \; a_3^i = \frac{(i+1)(i+2)}{2}[\kappa(\theta)]^i \kappa'(\theta), \; i \in \mathbb{N}_0, \tag{2.13}$$

$$a_1^i = a_2^i = a_3^i = 0 \qquad \text{for } i < 0. \tag{2.14}$$

We note that (2.11) in fact converges for $|\varepsilon\widehat{\rho}\kappa(\theta)| < 1$; this observation will be essential in our error estimates for the remainder in section 2.4.4.

Now, in order to define the inner expansion, we make the formal ansatz $u^{BL} = \sum_{i=0}^{\infty} \varepsilon^i \widehat{U}_i(\widehat{\rho}, \theta)$, where the functions $\widehat{U}_i$ are to be determined. Setting $L_\varepsilon u^{BL} = 0$ in (2.11) yields

$$\sum_{i=0}^{\infty} \varepsilon^i \sum_{j=0}^{i} L_j \widehat{U}_{i-j} = 0.$$

Hence, upon setting the coefficients of this formal power series in ε to zero, we obtain a recurrence relation for the sought functions $\widehat{U}_i$:

$$-\partial_{\widehat{\rho}}^2 \widehat{U}_i + \widehat{U}_i = \widehat{F}_i = \widehat{F}_i^1 + \widehat{F}_i^2 + \widehat{F}_i^3, \qquad i = 0, 1, \ldots, \tag{2.15}$$

$$\widehat{F}_i^1 = \sum_{j=0}^{i-1} \widehat{\rho}^j a_1^j \partial_{\widehat{\rho}} \widehat{U}_{i-1-j},\ \widehat{F}_i^2 = \sum_{j=0}^{i-2} \widehat{\rho}^j a_2^j \partial_\theta^2 \widehat{U}_{i-2-j},\ \widehat{F}_i^3 = \sum_{j=0}^{i-3} \widehat{\rho}^{j+1} a_3^j \partial_\theta \widehat{U}_{i-3-j}, \tag{2.16}$$

where we used the tacit convention that empty sums take the value zero. As we expect the boundary layer function u^{BL} to decay away from the boundary $\partial\Omega$ and as we want to satisfy the boundary conditions, we supplement these ODEs for the $\widehat{U}_i$ with the boundary conditions

$$\widehat{U}_i \to 0 \qquad \text{as } \widehat{\rho} \to \infty, \tag{2.17}$$

$$[\widehat{U}_i]_{\partial\Omega} = G_i : \begin{cases} = & g - [f]_{\partial\Omega} & \text{if } i = 0, \\ = & -[\Delta^{(i/2)} f]_{\partial\Omega} & \text{if } i \in \mathbb{N} \text{ is even}, \\ & 0 & \text{if } i \in \mathbb{N} \text{ is odd}. \end{cases} \tag{2.18}$$

The inner expansion of order $2M+1$ is defined as the function

$$u_M^{BL}(\rho, \theta) := \sum_{i=0}^{2M+1} \varepsilon^i \widehat{U}_i(\widehat{\rho}, \theta) = \sum_{i=0}^{2M+1} \varepsilon^i \widehat{U}_i(\rho/\varepsilon, \theta), \tag{2.19}$$

and it satisfies the desired boundary conditions

$$[u_M^{BL}]_{\partial\Omega} = g - \sum_{i=0}^{M} \varepsilon^{2i} [\Delta^{(i)} f]_{\partial\Omega}.$$

Remark 2.1. We defined u_M^{BL} as the inner expansion of order $2M+1$ so that the first neglected term of the formal asymptotic expansion $\sum_{i=0}^{\infty} \varepsilon^i \widehat{U}_i$ is of order ε^{2M+2}. This is precisely the same power of ε as the first neglected term of the outer expansion $\sum_{i=0}^{\infty} \varepsilon^{2i} \Delta^{(i)} f$ truncated after the ε^{2M} term.

Finally, the remainder r_M is defined such that (2.1) holds. We should note, however, that the boundary layer function u_M^{BL} and the cutoff function χ are defined in boundary fitted coordinates, whereas w_M is defined in the usual x, y coordinates so that, strictly speaking, the term χu_M^{BL} has to be understood as $(\chi u_M^{BL}) \circ \psi$ on the half-tubular neighborhood Ω_0 where ψ is the boundary fitted coordinate transformation defined in (2.3) and χu_M^{BL} is understood to vanish outside Ω_0.

2.3. Analytic regularity results for the asymptotic expansion. For every $M \in \mathbb{N}_0$ and cutoff function χ as in section 2.1 we can decompose the solution u_ε of (1.1) as in (2.1) where the outer expansion w_M, the inner expansion u_M^{BL}, and the remainder r_M are defined in the preceding section. The following theorem contains

bounds for w_M, u_M^{BL}, and r_M which are explicit in ε and M; the proof of this theorem is relegated to section 2.4.

THEOREM 2.2. *Let f, g, $\partial\Omega$ be analytic and let ρ_0, χ satisfy* (2.2), (2.5). *Let w_M, u_M^{BL}, r_M be defined by* (2.7), (2.19), (2.1). *Then there are constants C, K_1, and $K_2 > 0$ depending only on f, g, $\partial\Omega$, and ρ_0, χ such that*

(i) *For every $M \in \mathbb{N}_0$ the outer expansion w_M is analytic on Ω and there holds*

$$\|D^\alpha w_M\|_{L^\infty(\Omega)} \le C K_1^{|\alpha|} |\alpha|! (1 + (\varepsilon\, 2M\, K_2)^{2M}) \qquad \forall \alpha \in \mathbb{N}_0^2.$$

(ii) *For every $M \in \mathbb{N}_0$ the inner expansion u_M^{BL} is analytic on $(0,\rho_0)\times\mathbb{T}_L$. For every $\alpha \in [0,1)$ and all p, $m \in \mathbb{N}_0$, $(\rho,\theta) \in (0,\rho_0)\times\mathbb{T}_L$ there holds*

$$\left|\partial_\rho^p \partial_\theta^m u_M^{BL}(\rho,\theta)\right| \le C \left(1 + \left(\frac{\varepsilon(2M+1)K_2}{1-\alpha}\right)^{2M+1}\right) m! K_1^{m+p} \varepsilon^{-p} e^{-\alpha\rho/\varepsilon}.$$

(iii) *For every $M \in \mathbb{N}_0$ the remainder satisfies $r_M = 0$ on $\partial\Omega$ and*

$$\|r_M\|_{\varepsilon,\Omega} \le C \left(\varepsilon K_2 (2M+2)\right)^{2M+2}.$$

Remark 2.3. We notice that the crucial quantity in all three theorems is the product $\varepsilon(2M+2)$: If $\varepsilon(2M+2) \le q < K_2^{-1}$ then the remainder r_M is indeed small and we obtain bounds on the growth of the derivatives of w_M and u_M^{BL} which are independent of ε and M. In the complementary case, i.e., when $\varepsilon(2M+2)$ is not small, the asymptotic expansions lose their meaning.

We conclude this section with a few remarks concerning estimates on r_M.

Remark 2.4. Inspection of the proof of Theorem 2.2 (iii) below shows that in fact the following, slightly stronger results can be obtained:

1. There are C, $K > 0$ independent of ε and M such that

$$\|r_M\|_{\varepsilon,\Omega} \le C\Big\{\varepsilon^{2M+2}\|\Delta^{(M+1)} f\|_{L^\infty(\Omega)} + \varepsilon^{1/2}\left(K\varepsilon(2M+2)\right)^{2M+2}\Big\}.$$

   Hence, if the right-hand side f satisfies $\Delta^{(M+1)} f = 0$, e.g., if f is a polynomial of degree $2M+1$, then the ε-dependence of the estimate is actually improved by a factor $\varepsilon^{1/2}$.
2. In the proof of Theorem 2.2 (iii), with the exception of $\Delta^{(M+1)} f$, all the terms could be bounded in exponentially weighted spaces. This means that if $\Delta^{(M+1)} f = 0$, then we have estimates of the form

$$\|e^{\beta d(x)/\varepsilon} L_\varepsilon r_M\|_{L^2(\Omega)} \le C(\Omega,\beta)\varepsilon^{1/2}\left(K\varepsilon(2M+2)\right)^{2M+2},$$

   where $d(x) = \operatorname{dist}(x,\partial\Omega)$ and $\beta > 0$ appropriately. From this, one could infer estimates on r_M in exponentially weighted energy norms.
3. As $\|L_\varepsilon r_M\|_{L^\infty(\Omega)} \le C\left(K\varepsilon(2M+2)\right)^{2M+2}$ and $r_M = 0$ on $\partial\Omega$, the classical maximum principle gives us the pointwise bound

$$\|r_M\|_{L^\infty(\Omega)} \le C\left(K\varepsilon(2M+2)\right)^{2M+2}.$$

As the boundary $\partial\Omega$ is smooth, we can actually use the shift theorem for $-\Delta$ in order to control higher derivatives of r_M.

COROLLARY 2.5. *Assume the same hypotheses as in Theorem* 2.2. *Then for each $k \in \mathbb{N}_0$ there are constants C_k, $K > 0$ depending only on k, f, g, $\partial\Omega$, χ such that*

$$\|r_M\|_{H^k(\Omega)} \le C_k \varepsilon^{-k}\left(\varepsilon K(2M+2)\right)^{2M+2}, \qquad k \in \mathbb{N}_0.$$

Proof. The proof is an application of the classical shift theorem and an induction argument on k. We note that the corollary holds true for $k = 0$ and $k = 1$ by Theorem 2.2 (iii). Furthermore, r_M solves

$$-\Delta r_M = \varepsilon^{-2} L_\varepsilon r_M - \varepsilon^{-2} r_M \quad \text{in } \Omega, \qquad r_M = 0 \quad \text{on } \partial\Omega. \tag{2.20}$$

If we proceed as in the proof of Theorem 2.2 (iii) below but use the bounds on higher derivatives in Lemma 2.13 we can estimate

$$\|L_\varepsilon u_M^{BL}\|_{H^{k-2}(\Omega)} \le C_k \varepsilon^{2-k} \left(\varepsilon K(2M+2)\right)^{2M+2}, \qquad k \ge 2.$$

Hence the shift theorem allows us to conclude

$$\|r_M\|_{H^k(\Omega)} \le C_k \left(\varepsilon^{-k} \left(\varepsilon K(2M+2)\right)^{2M+2} + \varepsilon^{-2} \|r_M\|_{H^{k-2}(\Omega)}\right)$$

for $k \ge 2$. The obvious induction argument concludes the proof. ∎

2.4. Proof of Theorem 2.2.

2.4.1. Preliminaries. For $z \in \mathbb{C}$ and $\delta > 0$ we will denote by $B_\delta(z)$ the (open) disc in the complex plane of radius δ around z; $\partial B_\delta(z)$ denotes the positively oriented circle of radius δ with center z.

As the functions κ, G_i, and a_l^i are analytic on $\mathbb{T}_L$ these functions have holomorphic extensions to some complex neighborhood $S(\Theta)$ of the real line where

$$S(\Theta) := \{\theta \in \mathbb{C} \,|\, |\mathrm{Im}\theta| < \Theta\}, \quad \Theta > 0. \tag{2.21}$$

For future reference, let us note the following.

LEMMA 2.6. *Let f, g, $\partial\Omega$ be analytic. Then there are $\Theta > 0$ and C_G, C_A, γ_G, $A > 0$ with $A\rho_0 < 1$ such that $\forall \theta \in S(\Theta)$ and $\forall i \in \mathbb{N}_0$*

$$|G_i(\theta)| \le C_G i^i \gamma_G^i, \qquad |\kappa(\theta)| \le A, \qquad |a_1^i(\theta)| + |a_2^i(\theta)| + |a_3^i(\theta)| \le C_A A^i.$$

Proof. By (2.2) we may choose an A such that $\|\kappa\|_{L^\infty(\mathbb{T}_L)} < A < 1/\rho_0$. The bounds on G_i follow immediately from (2.6) and (2.18) for Θ sufficiently small. As $\rho_0 \|\kappa\|_{L^\infty(\mathbb{T}_L)} < 1$ the bound on κ follows from a continuity argument if Θ is sufficiently small. From this bound on κ, the bounds on the coefficients a_l^i can be obtained in view of (2.13) if Θ is taken sufficiently small and A is slightly enlarged. ∎

The following two lemmas can be proven by elementary considerations.

LEMMA 2.7. *$\forall\, q \ge 0$ and $\forall\, M \in \mathbb{N}_0$ there holds*

$$\sum_{i=0}^{M} q^i \le 2(1 + (4q)^M).$$

LEMMA 2.8. *Let $\alpha \in [0,1)$, $a \ge 0$, $M \ge 0$. Then*

$$\sup_{r \ge 0} \left| (M + a + r)^M e^{-(1-\alpha)r} \right| \le M^M (1-\alpha)^{-M} e^{(1-\alpha)a}.$$

2.4.2. Proof of Theorem 2.2 (i). In view of (2.6) the function w_M can be extended to a holomorphic function on a neighborhood $\tilde{\Omega} \subset \mathbb{C} \times \mathbb{C}$ of Ω which is independent of ε and M. Cauchy's theorem for derivatives allows us to infer the existence of C, $K > 0$ such that for each $(x, y) \in \Omega$ and all $\alpha \in \mathbb{N}_0^2$ there holds

$$|D^\alpha w_M(x,y)| \le CK^{|\alpha|}|\alpha|! \sum_{i=0}^{M} \varepsilon^{2i} \|\Delta^{(i)} f\|_{L^\infty(\tilde{\Omega})} \le CK^{|\alpha|}|\alpha|! \sum_{i=0}^{M} \gamma^{2i} (2i\varepsilon)^{2i}.$$

Estimating $\gamma(2i\varepsilon) \le \gamma(2M\varepsilon)$ we can easily obtain the desired result with the aid of Lemma 2.7.

2.4.3. Proof of Theorem 2.2 (ii). The proof of Theorem 2.2 (ii) is based on getting sharp bounds on the functions $\widehat{U}_i$ defined in (2.15)–(2.18). As the functions $\widehat{U}_i$ are solutions of ODEs whose right-hand sides depend on derivatives of the $\widehat{U}_j$, $0 \le j < i$, the following two lemmas will be necessary.

LEMMA 2.9. *Let f be an entire function satisfying for some $C_f > 0$, $j \in \mathbb{N}_0$*

$$|f(z)| \le C_f e^{-\mathrm{Re} z} (2 + j + |z|)^j \qquad \forall z \in \mathbb{C}.$$

Let $g \in \mathbb{C}$ and let $u : (0, \infty) \to \mathbb{C}$ be the solution of

$$-u'' + u = f \quad \text{on } (0,\infty), \qquad u(0) = g, \quad \lim_{x\to\infty} u(x) = 0.$$

Then u can be extended to an entire function (again denoted u) which satisfies

$$|u(z)| \le \left[C_f (2 + j + |z|)^{j+1} (j+1)^{-1} + |g| \right] e^{-\mathrm{Re} z} \qquad \forall z \in \mathbb{C}.$$

Proof. For $z \in (0, \infty)$, the use of a Green's function gives the following representation of the solution $u(z)$:

$$u(z) = \frac{1}{2} \left(e^{-z} \int_0^z e^y f(y)\, dy + e^z \int_z^\infty e^{-y} f(y)\, dy - e^{-z} \int_0^\infty e^{-y} f(y)\, dy \right) + g e^{-z}.$$

Analytic continuation then removes the restriction to $(0, \infty)$. In order to get the desired bound, we estimate each of these four terms separately. For the first integral, we use as the path of integration the straight line connecting 0 and z to get

$$\begin{aligned} \left| e^{-z} \int_0^z e^y f(y)\, dy \right| &\le e^{-\mathrm{Re} z} \int_0^1 C_f (2 + j + t|z|)^j |z|\, e^{-\mathrm{Re} tz} e^{\mathrm{Re} tz}\, dt \\ &\le C_f e^{-\mathrm{Re} z} \frac{1}{j+1} \left((2 + j + |z|)^{j+1} - (2+j)^{j+1} \right). \end{aligned}$$

For the second term, we calculate

$$\begin{aligned} \left| e^z \int_z^\infty e^{-y} f(y)\, dy \right| = \left| \int_0^\infty e^{-y} f(z+y)\, dy \right| &\le e^{-\mathrm{Re} z} C_f \int_0^\infty e^{-2y} (2 + j + |z| + y)^j\, dy \\ &= C_f e^{-\mathrm{Re} z} 2^{-(j+1)} e^{2(2+j+|z|)} \Gamma(j+1, 2(2+j+|z|)), \end{aligned}$$

where $\Gamma(\cdot,\cdot)$ denotes the incomplete Gamma function and we used Eq. 8.353.5 of [10] in the last step. Employing the estimate

$$|\Gamma(\alpha, \xi)| \le \frac{\left| e^{-\xi} \xi^\alpha \right|}{|\xi| - \alpha_0}, \qquad \alpha_0 = \max\{\alpha - 1, 0\}, \quad \mathrm{Re}\,\xi \ge 0, \quad |\xi| > \alpha_0$$

(see, e.g., Chap. 4, Sec. 10 of [11]) we finally arrive at

$$\left|e^z \int_z^\infty e^{-y} f(y)\, dy\right| \le C_f e^{-\mathrm{Re}z} \frac{(2+j+|z|)^{j+1}}{4+j+2|z|} \le C_f e^{-\mathrm{Re}z} \frac{(2+j+|z|)^{j+1}}{1+j}.$$

For the third term, we observe that the integral $\int_0^\infty f(y)e^{-y}\, dy$ is precisely the second term with $z = 0$. We conclude therefore that for the third term

$$\left|e^{-z} \int_0^\infty f(y)e^{-y}\, dy\right| \le C_f e^{-\mathrm{Re}z} \frac{(2+j)^{j+1}}{1+j}.$$

Combining these three estimates with the obvious one for the fourth term, we arrive at the desired bound. ∎

LEMMA 2.10. *Let U be holomorphic on $\mathbb{C} \times S(\Theta)$ for some $\Theta > 0$ and assume that there are $C_U > 0$, $i \in \mathbb{N}_0$ such that for all $\delta \in (0, \Theta)$*

$$\forall (z,\theta) \in \mathbb{C} \times S(\Theta - \delta) \qquad |U(z,\theta)| \le C_U (1+i+|z|)^i e^{-\mathrm{Re}z} \delta^{-i}.$$

Then $\forall\ \delta \in (0,\Theta)$ and $\forall\ (z,\theta) \in \mathbb{C} \times S(\Theta - \delta)$

$$\begin{aligned}
&|\partial_z U(z,\theta)| \le e\, C_U (2+i+|z|)^i e^{-\mathrm{Re}\ z} \delta^{-i},\\
&|\partial_\theta U(z,\theta)| \le e(i+1) C_U (1+i+|z|)^i e^{-\mathrm{Re}\ z} \delta^{-(i+1)},\\
&\left|\partial_\theta^2 U(z,\theta)\right| \le 2e(i+1)(i+2) C_U (1+i+|z|)^i e^{-\mathrm{Re}\ z} \delta^{-(i+2)}.
\end{aligned}$$

Proof. For the first estimate, we use Cauchy's integral theorem for derivatives:

$$|\partial_z U(z,\theta)| = \left|\frac{1}{2\pi i} \int_{|t|=1} \frac{U(z+t,\theta)}{t^2}\, dt\right| \le C_U (1+i+|z|+1)^i e^{-\mathrm{Re}z+1} \delta^{-i}.$$

For the second and third estimate, we again use Cauchy's integral theorem for derivatives but with the path of integration chosen as $\partial B_{\kappa\delta}(\theta)$ where $\kappa \in (0,1)$ is to be chosen below. For the second estimate, we arrive at

$$\begin{aligned}
|\partial_\theta U(z,\theta)| &= \left|\frac{1}{2\pi i} \int_{|t|=\kappa\delta} \frac{U(z,\theta+t)}{t^2}\, dt\right| \le C_U \frac{1}{(\kappa\delta)((1-\kappa)\delta)^i} (1+i+|z|)^i e^{-\mathrm{Re}z}\\
&\le C_U \frac{1}{\kappa(1-\kappa)^i} \delta^{-(i+1)} (1+i+|z|)^i e^{-\mathrm{Re}z}.
\end{aligned}$$

Choosing $\kappa = 1/(i+2)$ and observing that with this choice $e^{-1}(i+1)^{-1} \le \kappa(1-\kappa)^i$ holds for all $i \in \mathbb{N}_0$ we obtain the desired second bound. Finally, the third estimate is proved completely analogously. ∎

These two lemmas put us in a position to obtain results about the functions $\widehat{U}_i$ defined by (2.15)–(2.18). As $\widehat{F}_0 = 0$ we have $\widehat{U}_0(\widehat{\rho},\theta) = G_0(\theta)e^{-\widehat{\rho}}$. Therefore, $\widehat{F}_1(\widehat{\rho},\theta) = -a_1^0(\theta)G_0(\theta)e^{-\widehat{\rho}}$ and we get $\widehat{U}_1(\widehat{\rho},\theta) = -1/2\widehat{\rho}a_1^0(\theta)e^{-\widehat{\rho}} + G_1(\theta)e^{-\widehat{\rho}}$. It is easy to see that in general the functions $\widehat{U}_i$ are of the form $e^{-\widehat{\rho}}$ times a polynomial of degree i in $\widehat{\rho}$ whose coefficients involve the functions a_l^j and G_j and their derivatives. Furthermore, the lowest-order term of the polynomial of degree i is $G_i(\theta)$ which by Lemma 2.6 can be estimated by $C_G i^i \gamma_G^i$ for θ in a neighborhood of the real line. This suggests that bounds on $\widehat{U}_i$ of the form $|\widehat{U}_i(\widehat{\rho},\theta)| \le CK^i e^{-\mathbf{Re}\widehat{\rho}}(1+i+|\widehat{\rho}|)^i$ can be obtained.

LEMMA 2.11. *Let f, g, $\partial\Omega$ be analytic. Then there are C_U, K, $\Theta > 0$ depending only on f, g, $\partial\Omega$ such that the functions $\widehat{U}_i$ defined by (2.15)–(2.18) are holomorphic on $\mathbb{C} \times S(\Theta)$ and satisfy $\forall\, i \in \mathbb{N}_0$*

$$\left|\widehat{U}_i(\widehat{\rho},\theta)\right| \le C_U K^i (1+i+|\widehat{\rho}|)^i e^{-\mathrm{Re}\,\widehat{\rho}} \qquad \forall (\widehat{\rho},\theta) \in \mathbb{C}\times S(\Theta).$$

Proof. We prove the following result: There are C_U, K, Θ such that for all $i \in \mathbb{N}_0$

(2.22)

$$\forall\, \delta \in (0,\Theta) \quad \forall (\widehat{\rho},\theta) \in \mathbb{C}\times S(\Theta-\delta) \qquad |\widehat{U}_i(\widehat{\rho},\theta)| \le C_U K^i \delta^{-i} (1+i+|\widehat{\rho}|)^i e^{-\mathrm{Re}\,\widehat{\rho}}.$$

The claim of the lemma then follows by slightly decreasing Θ and adjusting the constant K. Let us assume that Θ, C_A, C_G, γ_G, A are chosen as in Lemma 2.6. From our preliminary discussion it is clear that we can (after possibly decreasing Θ slightly) choose C_U, K such that (2.22) is satisfied for $i = 0, 1, 2$. Let us furthermore assume that C_U and K are chosen so large such that $K > \max\{A\Theta, \gamma_G\Theta\}$ and that $\forall\, i \in \mathbb{N}_0$

$$\left[K^{-1}\frac{eC_A\Theta}{1-A\Theta/K} + K^{-2}\frac{2eC_A}{1-A\Theta/K} + K^{-3}\frac{eC_A\Theta^2}{1-A\Theta/K} + \frac{C_G}{C_U}\left(\frac{\gamma_G\Theta}{K}\right)^i\right] \le 1. \tag{2.23}$$

In order to proceed by induction on i, let $i \ge 3$ and assume that (2.22) holds for all $0 \le j \le i-1$. To get the desired bounds on $\widehat{U}_i$ by means of Lemma 2.9 we need to control $\widehat{F}_i^1$, $\widehat{F}_i^2$, $\widehat{F}_i^3$. Combining Lemma 2.10 and the induction hypothesis (2.22), we get for $(\widehat{\rho},\theta) \in \mathbb{C}\times S(\Theta-\delta)$

$$\begin{aligned}
|\widehat{F}_i^1(\widehat{\rho},\theta)| &\le \sum_{j=0}^{i-1} |\widehat{\rho}|^j |a_1^j(\theta)||\partial_{\widehat{\rho}}\widehat{U}_{i-1-j}(\widehat{\rho},\theta)| \\
&\le C_U K^i \delta^{-i} e^{-\mathrm{Re}\,\widehat{\rho}} C_A K^{-1}\delta e \sum_{j=0}^{i-1} |\widehat{\rho}|^j A^j K^{-j}\delta^j (i-j+1+|\widehat{\rho}|)^{i-1-j} \\
&\le C_U K^i \delta^{-i} K^{-1} \frac{eC_A\delta}{1-A\delta/K}(i+1+|\widehat{\rho}|)^{i-1} e^{-\mathrm{Re}\,\widehat{\rho}}.
\end{aligned}$$

Similarly, we obtain for $|\widehat{F}_i^2|$ and $|\widehat{F}_i^3|$ on $\mathbb{C}\times S(\Theta-\delta)$

$$\begin{aligned}
|\widehat{F}_i^2(\widehat{\rho},\theta)| &\le C_U K^i \delta^{-i} K^{-2} \frac{2eC_A}{1-A\delta/K}(i+|\widehat{\rho}|)^{i-2}(i-1) i e^{-\mathrm{Re}\,\widehat{\rho}}, \\
|\widehat{F}_i^3(\widehat{\rho},\theta)| &\le C_U K^i \delta^{-i} K^{-3} \frac{eC_A\delta^2}{1-A\delta/K}(i+|\widehat{\rho}|)^{i-2}(i-2) e^{-\mathrm{Re}\,\widehat{\rho}}.
\end{aligned}$$

Hence, applying Lemma 2.9 with $\widehat{F}_i^1$, $\widehat{F}_i^2$, $\widehat{F}_i^3$ as right-hand sides f (and initial condition $g = 0$) and finally with homogeneous right-hand side and initial condition $g = G_i(\theta)$, we obtain

$$\begin{aligned}
|\widehat{U}_i(\widehat{\rho},\theta)| \le C_U K^i \delta^{-i} e^{-\mathbf{Re}\widehat{\rho}} \Bigg[& K^{-1}\frac{eC_A\delta}{1-A\delta/K}\frac{(i+1+|\widehat{\rho}|)^i}{i} \\
& + K^{-2}\frac{2eC_A}{1-A\delta/K}(i+|\widehat{\rho}|)^{i-1}\frac{(i-1)i}{i-1}
\end{aligned}$$

$$
\begin{aligned}
&+K^{-3}\frac{eC_A\delta^2}{1-A\delta/K}(i+|\widehat{\rho}|)^{i-1}\frac{(i-2)}{i-1}+\frac{C_G}{C_U}\left(\frac{\delta\gamma_G}{K}\right)^i\Bigg]\\
&\le C_UK^i\delta^{-i}e^{-\mathbf{Re}\widehat{\rho}}(i+1+|\widehat{\rho}|)^i\times\\
&\quad\times\left[K^{-1}\frac{eC_A\delta}{1-A\delta/K}+K^{-2}\frac{2eC_A}{1-A\delta/K}+K^{-3}\frac{eC_A\delta^2}{1-A\delta/K}+\frac{C_G}{C_U}\left(\frac{\delta\gamma_G}{K}\right)^i\right]\\
&\le C_UK^i\delta^{-i}e^{-\mathbf{Re}\widehat{\rho}}(i+1+|\widehat{\rho}|)^i
\end{aligned}
$$

as the bracketed expression is bounded by 1 by the choice of K in (2.23). This concludes the induction argument. ∎

The proof of Theorem 2.2 (ii) is now straightforward.

Proof of Theorem 2.2 (ii). Let $\alpha\in[0,1)$ be given. By Lemma 2.11 the functions $\widehat{U}_i$ are holomorphic on $\mathbb{C}\times S(\Theta)$ for some $\Theta>0$. We obtain with Cauchy's integral formula for derivatives for p, $m\in\mathbb{N}_0$ and $\rho\ge0$, $\theta\in\mathbb{T}_L$

$$
\partial_\rho^p\,\partial_\theta^m\,\widehat{U}_i(\rho/\varepsilon,\theta)=-\varepsilon^{-p}\frac{p!m!}{4\pi^2}\int_{|z|=R}\int_{|t|=\Theta/2}\frac{\widehat{U}_i(\rho/\varepsilon+z,\theta+t)}{z^{p+1}t^{m+1}}\,dzdt.
$$

Now choosing $R=p+1$ and using Lemma 2.11 (note that we may choose $\delta=\Theta/2$), we get the existence of constants C, $K>0$ such that

$$
\left|\partial_\rho^p\,\partial_\theta^m\,\widehat{U}_i(\rho/\varepsilon,\theta)\right|=C\varepsilon^{-p}\frac{p!m!}{(p+1)^p}(2/\Theta)^mK^i(1+i+\rho/\varepsilon+p+1)^ie^{-\rho/\varepsilon+(p+1)}.
$$

Upon writing $e^{-\rho/\varepsilon}=e^{-(1-\alpha)\rho/\varepsilon}e^{-\alpha\rho/\varepsilon}$ and appealing to Lemma 2.8 (with $a=2+p$), we get, together with the estimate $p!\le Cp^pe^{-p}\sqrt{2\pi(p+1)}$ from Stirling's formula,

$$
\left|\partial_\rho^p\,\partial_\theta^m\,\widehat{U}_i(\rho/\varepsilon,\theta)\right|=C\varepsilon^{-p}e^{(1-\alpha)p}(p+1)^{1/2}m!(2/\Theta)^mK^ii^i(1-\alpha)^{-i}e^{-\alpha\rho/\varepsilon}. \tag{2.24}
$$

As $u_M^{BL}(\rho,\theta)=\sum_{i=0}^{2M+1}\varepsilon^i\widehat{U}_i(\rho/\varepsilon,\theta)$ estimating $i^i\le(2M+1)^i$ in (2.24) and summing gives the desired result after appealing to Lemma 2.7. ∎

2.4.4. Proof of Theorem 2.2 (iii). We start with a lemma concerning $L_\varepsilon u_M^{BL}$.

LEMMA 2.12. *Let u_M^{BL} be defined in* (2.19). *Then there are Θ, C, $K>0$ depending only on f, g, $\partial\Omega$ such that $L_\varepsilon u_M^{BL}$ is holomorphic on $B_{\rho_0}(0)\times S(\Theta)\subset\mathbb{C}\times\mathbb{C}$ and satisfies*

$$
|L_\varepsilon u_M^{BL}(\rho,\theta)|\le CK^{2M+2}(\varepsilon(2M+2)+|\rho|)^{2M+2}e^{-\mathbf{Re}\rho/\varepsilon}\qquad\forall(\rho,\theta)\in B_{\rho_0}(0)\times S(\Theta).
$$

Proof. As we observed above, the power series expansion (2.11) of L_ε converges absolutely for $|\varepsilon\widehat{\rho}\kappa(\theta)|<1$. In order to exploit this fact, let us choose with the aid of Lemma 2.6 $\Theta>0$, C_A, $A>0$ with $\rho_0A=:q<1$ such that $|\rho\kappa(\theta)|\le q<1$ $\forall$ $(\rho,\theta)\in B_{\rho_0}(0)\times S(\Theta)$. Furthermore, we may assume without loss of generality that Θ is so small that Lemma 2.11 holds.

By the construction of the functions $\widehat{U}_i$, we calculate directly (we write $\widehat{\rho}=\rho/\varepsilon$ whenever notationally convenient)

$$
L_\varepsilon u_M^{BL}(\rho,\theta)=\sum_{i=2M+2}^{\infty}\varepsilon^i\sum_{j=0}^{2M+1}L_{i-j}\widehat{U}_j(\rho/\varepsilon,\theta)
$$

$$= -\sum_{j=0}^{2M+1}\sum_{i=2M+2}^{\infty}\varepsilon^i\widehat{\rho}^{i-1-j}a_1^{i-1-j}\partial_{\widehat{\rho}}\widehat{U}_j - \sum_{j=0}^{2M+1}\sum_{i=2M+3}^{\infty}\varepsilon^i\widehat{\rho}^{i-2-j}a_2^{i-2-j}\partial_\theta^2\widehat{U}_j$$
$$- \sum_{j=0}^{2M+1}\sum_{i=2M+4}^{\infty}\varepsilon^i\widehat{\rho}^{i-2-j}a_3^{i-3-j}\partial_\theta\widehat{U}_j.$$

Each of these three terms can be estimated with the aid of Lemmas 2.11 and 2.10. For the first term, we have for $(\rho,\theta)\in B_{\rho_0}(0)\times S(\Theta)$ (which implies $|\rho|A\le q<1$)

$$\begin{aligned}
&\left|\sum_{j=0}^{2M+1}\sum_{i=2M+2}^{\infty}\varepsilon^i\widehat{\rho}^{i-1-j}a_1^{i-1-j}(\theta)\partial_{\widehat{\rho}}\widehat{U}_j(\widehat{\rho},\theta)\right|\\
&\le eC_UC_A\sum_{j=0}^{2M+1}\sum_{i=2M+2}^{\infty}\varepsilon^i|\widehat{\rho}|^{i-1-j}A^{i-1-j}K^j(2+j+|\widehat{\rho}|)^je^{-\mathrm{Re}\,\widehat{\rho}}\\
&\le eC_AC_U\varepsilon^{2M+2}\sum_{j=0}^{2M+1}\left(\sum_{i=0}^{\infty}\varepsilon^i|\widehat{\rho}|^iA^i\right)A^{2M+1-j}K^j\left\{|\widehat{\rho}|^{2M+1-j}(2M+3+|\widehat{\rho}|)^j\right\}e^{-\mathrm{Re}\widehat{\rho}}\\
&\le eC_AC_U\varepsilon^{2M+2}\frac{1}{1-q}\sum_{j=0}^{2M+1}A^{2M+1-j}K^j(2M+3+|\widehat{\rho}|)^{2M+1}e^{-\mathrm{Re}\,\widehat{\rho}}\\
&\le 2eC_AC_U\frac{1}{1-q}A^{2M+1}(1+(4K/A)^{2M+1})(\varepsilon(2M+3)+|\rho|)^{2M+2}e^{-\mathrm{Re}\,\widehat{\rho}},
\end{aligned}$$

where we exploited the assumption that $|\rho|A\le q<1$ and appealed to Lemma 2.7. The factor $(2M+3)$ can easily be replaced by $2M+2$ at the expense of a larger constant. Hence, the first sum leads to an estimate of the desired form; the remaining two sums are estimated similarly. ∎

A procedure similar to that of Theorem 2.2 (ii) gives bounds on higher derivatives.

LEMMA 2.13. *Let $\rho'<\rho_0$, $\alpha\in[0,1)$. Then there are constants C, K_1, $K_2>0$ independent of ε, M such that for all p, $m\in\mathbb{N}_0$ and all $(\rho,\theta)\in[0,\rho')\times\mathbb{T}_L$*

$$\left|\partial_\rho^p\partial_\theta^m L_\varepsilon u_M^{BL}(\rho,\theta)\right|\le C\varepsilon^{-p}p!m!K_1^{p+m}\left(\frac{\varepsilon(2M+2)K_2}{1-\alpha}\right)^{2M+2}e^{-\alpha\rho/\varepsilon}.$$

Proof. As $\rho'<\rho_0$ there is $\delta>0$ such that $B_{\delta\varepsilon}(\rho)\subset B_{\rho_0}(0)$ for all $\varepsilon\in(0,1]$ and $\rho\in[0,\rho')$. We apply Cauchy's integral theorem for derivatives to calculate $\partial_\rho^p\partial_\theta^m L_\varepsilon u_M^{BL}$; the path of integration is chosen as $\partial B_{\delta\varepsilon}(\rho)$ for the first variable and as $\partial B_{\Theta/2}(\theta)$ for the second variable. Together with Lemma 2.12 we get the existence of C, K_1, $K_2>0$ such that for $\rho\in[0,\rho')$, $\theta\in\mathbb{T}_L$

$$\begin{aligned}
\left|\partial_\rho^p\partial_\theta^m L_\varepsilon u_M^{BL}(\rho,\theta)\right| &\le C\varepsilon^{-p}p!m!K_1^{p+m}K_2^{2M+2}\left(\varepsilon(2M+2)+\rho+\varepsilon\delta\right)^{2M+2}e^{-\rho/\varepsilon}\\
&\le C\varepsilon^{-p}p!m!K_1^{p+m}K_2^{2M+2}\varepsilon^{2M+2}\left[\left((2M+2)+\rho/\varepsilon+\delta\right)^{2M+2}e^{-(1-\alpha)\rho/\varepsilon}\right]e^{-\alpha\rho/\varepsilon}.
\end{aligned}$$

Appealing to Lemma 2.8 to bound the expression in brackets allows us to conclude the argument. ∎

We are now in a position to estimate the accuracy of the asymptotic expansion, i.e., bound r_M.

Proof of Theorem 2.2 (iii). For any $M \in \mathbb{N}_0$ the remainder r_M is defined as $r_M = u_\varepsilon - w_M - \chi u_M^{BL}$, where w_M is defined by (2.7), u_M^{BL} is defined by (2.19), and χ is the cutoff function of (2.5). Hence, by construction of u_M^{BL}, $r_M = 0$ on $\partial\Omega$. Furthermore, the remainder r_M solves the following elliptic equation:

$$\begin{aligned} L_\varepsilon r_M &= L_\varepsilon \left(u_\varepsilon - w_M - \chi u_M^{BL}\right) = \varepsilon^{2M+2}\Delta^{(M+1)} f - L_\varepsilon(\chi u_M^{BL}) \\ &= \varepsilon^{2M+2}\Delta^{(M+1)} f + \varepsilon^2(\Delta\chi)u_M^{BL} + 2\varepsilon^2 \nabla\chi\cdot\nabla u_M^{BL} - \chi L_\varepsilon u_M^{BL}. \end{aligned}$$

From (1.3) we infer

$$\|r_M\|_{\varepsilon,\Omega} \le \|L_\varepsilon r_M\|_{L^2(\Omega)}. \tag{2.25}$$

We are therefore left with estimating the L^2 norm of the four terms that add up to $L_\varepsilon r_M$. By the assumptions on f, cf. (2.6), we have

$$\|\varepsilon^{2M+2}\Delta^{(M+1)} f\|_{L^\infty(\Omega)} \le C_f \left(\varepsilon\gamma(2M+2)\right)^{2M+2}. \tag{2.26}$$

Let us fix $\alpha \in (0,1)$ for the remainder of this proof. As $\chi \equiv 1$ for $0 < \rho < \rho_1$ and $\chi \equiv 0$ for $\rho > (\rho_1+\rho_0)/2$, we obtain with the aid of Theorem 2.2 (ii)

$$\begin{aligned} \varepsilon^2\|(\Delta\chi)u_M^{BL}\|_{L^2(\Omega)} &\le C\varepsilon^2 S_M e^{-\alpha\rho_1/\varepsilon}, \\ \varepsilon^2\|\nabla\chi\cdot\nabla u_M^{BL}\|_{L^2(\Omega)} &\le C\varepsilon S_M e^{-\alpha\rho_1/\varepsilon}, \end{aligned}$$

where $S_M = 1 + (\varepsilon(2M+1)K)^{2M+1}$ for some appropriate C, $K > 0$. We calculate further

$$\begin{aligned} S_M &= 1 + (\varepsilon(2M+1)K)^{2M+1} \le (1+\varepsilon(2M+2)K)^{2M+2} \\ &\le \varepsilon^{2M+2}K^{2M+2}(2M+2+K^{-1}/\varepsilon)^{2M+2}. \end{aligned}$$

Without loss of generality, $K^{-1} \le \rho_1$ and therefore we obtain by appealing to Lemma 2.8

$$\varepsilon S_M e^{-\alpha\rho_1/\varepsilon} \le \varepsilon\varepsilon^{2M+2}K^{2M+2}(2M+2+\rho_1/\varepsilon)^{2M+2}e^{-\alpha\rho_1/\varepsilon} \le \varepsilon(\varepsilon(2M+2)K\alpha^{-1})^{2M+2}.$$

Finally, exploiting the exponential decay normal to $\partial\Omega$, it is easy to deduce from Lemma 2.13 that

$$\|\chi L_\varepsilon u_M^{BL}\|_{L^2(\Omega)} \le C\varepsilon^{1/2}\left(K'\varepsilon(2M+2)\right)^{2M+2}$$

for some $K' > 0$ independent of ε and M. Inserting all these estimates into (2.25), we obtain for some $C > 0$ that

$$\begin{aligned} \|r_M\|_{\varepsilon,\Omega} \le C\Big(&(\varepsilon(2M+2)\gamma)^{2M+2} \\ &+ \varepsilon(\varepsilon(2M+2)K\alpha^{-1})^{2M+2} + \varepsilon^{1/2}(\varepsilon(2M+2)K')^{2M+2}\Big). \end{aligned}$$

As $\alpha \in (0,1)$ is fixed, the desired bound follows. ∎

3. Growth estimates for the derivatives. The main result of this section is the following bound on the growth of the derivatives of the solutions u_ε of (1.1).

THEOREM 3.1. *Let f, g, $\partial\Omega$ be analytic and let u_ε be the solution of* (1.1). *Then there are C and $K > 0$ depending only on f, g, and the geometry of Ω (in particular, C, K are independent of ε) such that*

$$\|D^\alpha u_\varepsilon\|_{L^2(\Omega)} \le CK^{|\alpha|}\max\left(|\alpha|,\varepsilon^{-1}\right)^{|\alpha|}\left(1+\|u_\varepsilon\|_{\varepsilon,\Omega}\right) \qquad \forall\alpha \in \mathbb{N}_0^2. \tag{3.1}$$

Remark 3.2. The proof of Theorem 3.1 below shows that (3.1) holds true for solutions u_ε of (1.1) with $\varepsilon \in \mathbb{C}$ (ε^{-1} has to be replaced with $|\varepsilon^{-1}|$ in (3.1)) and hence the case of Helmholtz's equation is also covered. Furthermore, (3.1) still holds true if the right-hand side f is allowed to have "boundary layer character" in the sense that there exist C, $K > 0$ such that

$$\|D^\alpha f\|_{L^2(\Omega)} \le CK^{|\alpha|} \max(|\alpha|, |\varepsilon^{-1}|)^{|\alpha|} \qquad \forall \alpha \in \mathbb{N}_0^2.$$

The remainder of this paper is devoted to the proof of Theorem 3.1. To that end, we will state and prove some local analytic regularity results for the solution u_ε in sections 3.1 and 3.2. The proof of Theorem 3.1 in section 3.3 then concludes the paper.

3.1. Local analytic regularity results. The proof of Theorem 3.1 rests on two local analytic regularity results: an interior result on discs (Proposition 3.3) and a boundary result on half-discs (Proposition 3.4). Due to their technical nature, the proofs of these results of this section are deferred to section 3.2.

Our local results are very similar to those of section 5.7 of [8] and we therefore use the same notation: For $r > 0$ we define discs B_r and half-discs G_r by

$$B_r := B_r(0) \subset \mathbb{R}^2, \qquad G_r := \{(x, y) \in B_r \,|\, y > 0\}.$$

Furthermore, for smooth functions u and $R > 0$ we introduce

$$|\nabla^p u(x)|^2 := \sum_{|\alpha|=p} \frac{|\alpha|!}{\alpha!} |D^\alpha u(x)|^2 = \sum_{\beta_1,\dots,\beta_p=1}^{2} |D^{\beta_1 \cdots \beta_p} u(x)|^2 \qquad \forall p \in \mathbb{N}_0, \tag{3.2}$$

$$[p] := \max(1, p) \qquad \forall p \in \mathbb{Z}, \tag{3.3}$$

$$N_{R,p}(u) := \frac{1}{[p]!} \sup_{R/2 \le r < R} (R - r)^{2+p} \|\nabla^{p+2} u\|_{L^2(B_r)}, \quad p \in \mathbb{N}_0 \cup \{-2, -1\}, \tag{3.4}$$

$$N'_{R,p,q}(u) := \frac{1}{[p+q]!} \sup_{R/2 \le r < R} (R - r)^{p+q+2} \|\partial_y^{q+2} \partial_x^p u\|_{L^2(G_r)}, \; p \ge 0, q \ge -2. \tag{3.5}$$

Then we have the following local results.

PROPOSITION 3.3. *Let $R \in (0, 1]$ and let u be a solution of*

$$-\varepsilon^2 \Delta u + bu = f \qquad \text{on } B_R, \tag{3.6}$$

where b and f are analytic on B_R and satisfy for some C_B, B, C_f, $\gamma > 0$

$$\|\nabla^p b\|_{L^\infty(B_R)} \le C_B B^p p! \qquad \forall p \in \mathbb{N}_0, \tag{3.7}$$

$$\|\nabla^p f\|_{L^2(B_R)} \le C_f \gamma^p [p^p R^{-p} + \max(p, \varepsilon^{-1})^p] \qquad \forall p \in \mathbb{N}_0. \tag{3.8}$$

Then there is $K > 0$ independent of ε and R such that for all $p \ge -2$

$$N_{R,p}(u) \le C_u K^{p+2} \frac{\max([p], R/\varepsilon)^{p+2}}{[p]!}, \tag{3.9}$$

$$C_u = \min(1, R/\varepsilon)\varepsilon \|\nabla u\|_{L^2(B_R)} + \|u\|_{L^2(B_R)} + C_f \min(1, (R/\varepsilon)^2). \tag{3.10}$$

PROPOSITION 3.4. *Let $R \in (0, 1]$ and let u be a solution of*

$$-\varepsilon^2 \Delta u + bu = f \quad \text{on } G_R, \qquad u = 0 \text{ on } \partial G_R \cap \{(x, y) \,|\, y = 0\}, \tag{3.11}$$

where b and f are analytic on B_R and satisfy for some C_B, B, C_f, $\gamma > 0$

$$\|\nabla^p b\|_{L^\infty(G_R)} \le C_B B^p p! \qquad \forall p \in \mathbb{N}_0, \tag{3.12}$$

$$\|\nabla^p f\|_{L^2(G_R)} \le C_f \gamma^p [p^p R^{-p} + \max(p, \varepsilon^{-1})^p] \qquad \forall p \in \mathbb{N}_0. \tag{3.13}$$

Then there are K_1, $K_2 > 0$ independent of ε and R such that $\forall p \ge 0$, $q \ge -2$

$$N'_{R,p,q}(u) \le C_u K_1^{p+2} K_2^{q+2} \frac{\max([p+q], R/\varepsilon)^{p+q+2}}{[p+q]!}, \tag{3.14}$$

$$C_u = \min(1, R/\varepsilon)\varepsilon \|\nabla u\|_{L^2(G_R)} + \|u\|_{L^2(G_R)} + C_f \min(1, (R/\varepsilon)^2). \tag{3.15}$$

Remark 3.5. The local results in Propositions 3.3 and 3.4 make much weaker assumptions on the right-hand side f than Theorem 3.1. The term $\max(p, \varepsilon^{-1})^p$ appearing in (3.8) and (3.13) indicates that right-hand sides with boundary layer character are admissible as right-hand sides. Furthermore, in Propositions 3.3 and 3.4 the dependence on the radius of the discs R is given explicitly; this explicit dependence could be used to obtain results similar to Theorem 3.1 for domains Ω with piecewise analytic boundary $\partial\Omega$.

Finally, we need the following.

LEMMA 3.6. *Let G, $G_1 \subset \mathbb{R}^2$ be bounded open sets. Assume that $g = (g_1, g_2) : \overline{G}_1 \to \mathbb{R}^2$ is analytic and injective on $\overline{G}_1$, $\det g' \ne 0$ on $\overline{G}_1$, and satisfies $g(G_1) \subset G$. Let $f : \overline{G} \to \mathbb{C}$ be analytic on $\overline{G}$ and assume that it satisfies for some ε, C_f, $\gamma > 0$*

$$\|\nabla^p f\|_{L^2(G)} \le C_f \gamma^p \max(p, \varepsilon^{-1})^p \qquad \forall p \in \mathbb{N}_0.$$

Then there are C, $K > 0$ depending only on C_f, γ, and the map g such that

$$\|\nabla^p (f \circ g)\|_{L^2(G_1)} \le C K^p \max(p, \varepsilon^{-1})^p \qquad \forall p \in \mathbb{N}_0.$$

3.2. Proof of local regularity results.

3.2.1. Interior estimates (proof of Proposition 3.3). In order to prove Proposition 3.3, it is convenient to introduce for smooth functions u the quantity

$$M_{R,p}(u) := \frac{1}{p!} \sup_{R/2 \le r < R} (R - r)^{2+p} \|\nabla^p u\|_{L^2(B_r)}, \qquad p \in \mathbb{N}_0.$$

From standard elliptic regularity theory one can infer the following.

LEMMA 3.7. *Let u solve $\Delta u = f$ on B_R. Then there is $C_1 > 0$ independent of u, R, f such that*

$$N_{R,p}(u) \le C_1 [M_{R,p}(f) + N_{R,p-1}(u) + N_{R,p-2}(u)] \qquad \forall p \in \mathbb{N}_0.$$

Proof. The proof can be found in [8, Lem. 5.7.3]. ∎

LEMMA 3.8. *Let b, u be analytic and assume that b satisfies* (3.7). *Then*

$$M_{r,p}(bu) \le C_b \sum_{q=0}^{p} \left(B\frac{R}{2}\right)^{p-q} \left(\frac{R}{2}\right)^2 \frac{[q-2]!}{q!} N_{R,q-2}(u).$$

Proof. From Leibniz's formula, we have (cf. Lemma 5.7.4 of [8])

$$|\nabla^p(bu)| \le \sum_{q=0}^{p} \binom{p}{q} |\nabla^{p-q} b|\, |\nabla^q u|.$$

This allows us to get

$$\begin{aligned} M_{R,p}(bu) &\le \frac{1}{p!} \sup_{R/2\le r<R} (R-r)^{p+2} \|\nabla^p(bu)\|_{L^2(B_r)} \\ &\le \frac{1}{p!} \sup_{R/2\le r<R} (R-r)^{p+2} \sum_{q=0}^{p} \binom{p}{q} \|\nabla^{p-q} b\|_{L^\infty(B_R)} \|\nabla^q u\|_{L^2(B_r)} \\ &\le C_b \sum_{q=0}^{p} \left(\frac{BR}{2}\right)^{p-q} \left(\frac{R}{2}\right)^2 \frac{[q-2]!}{q!} N_{R,q-2}(u), \end{aligned}$$

which concludes the proof. ∎

Proof of Proposition 3.3. Let C_1 be the generic constant of Lemma 3.7 and choose $2K > \max(2, \gamma, BR)$ such that

$$C_1 \frac{1}{2} K^{-2} \left(\frac{\gamma R}{2K}\right)^p + C_1 \left(\frac{C_b/4}{1 - BR/(2K)} K^{-2} + K^{-1} + K^{-2} \right) \le 1 \qquad \forall p \in \mathbb{N}_0. \tag{3.16}$$

We will proceed by induction on p. As $K \ge 1$, the claim (3.9) holds for $p = -2$ and $p = -1$. Let us therefore assume that (3.9) holds $\forall\ -2 \le p' < p$. As $-\Delta u = \varepsilon^{-2}(f - bu)$, we get for $p \in \mathbb{N}_0$ using Lemmas 3.7 and 3.8

$$\begin{aligned} N_{R,p}(u) &\le C_1 \left\{ \varepsilon^{-2} M_{R,p}(f - bu) + N_{R,p-1}(u) + N_{R,p-2}(u) \right\} \\ &\le C_1 \Big\{ \varepsilon^{-2} M_{R,p}(f) + \varepsilon^{-2} C_b \sum_{q=0}^{p} \left(\frac{BR}{2}\right)^{p-q} \left(\frac{R}{2}\right)^2 \frac{[q-2]!}{q!} N_{R,q-2}(u) \\ &\qquad + N_{R,p-1}(u) + N_{R,p-2}(u) \Big\}. \end{aligned}$$

From the induction hypothesis (3.9) we obtain

$$\begin{aligned} N_{R,p}(u) \le C_1 \varepsilon^{-2} M_{R,p}(f) + C_1 C_u \Big\{ C_b \sum_{q=0}^{p} \left(\frac{BR}{2}\right)^{p-q} \left(\frac{R}{2}\right)^2 \varepsilon^{-2} K^q \frac{\max([q-2], R/\varepsilon)^q}{q!} \\ + K^{p+1} \frac{\max([p-1], R/\varepsilon)^{p+1}}{[p-1]!} + K^p \frac{\max([p-2], R/\varepsilon)^p}{[p-2]!} \Big\}. \end{aligned}$$

As we have the estimates

$$\begin{aligned} R^2 \varepsilon^{-2} \frac{1}{p!} \frac{p!}{q!} \max([q-2], R/\varepsilon)^q &\le \frac{1}{p!} \max([p], R/\varepsilon)^{p+2}, \\ \frac{1}{p!} \frac{p!}{[p-1]!} \max([p-1], R/\varepsilon)^{p+1} &\le \frac{1}{p!} \max([p], R/\varepsilon)^{p+2}, \\ \frac{1}{p!} \frac{p!}{[p-2]!} \max([p-2], R/\varepsilon)^{p} &\le \frac{1}{p!} \max([p], R/\varepsilon)^{p+2}, \end{aligned}$$

we obtain

$$\begin{aligned} N_{R,p}(u) &\le C_1\varepsilon^{-2}M_{R,p}(f) + \frac{\max([p],R/\varepsilon)^{p+2}}{p!}K^{p+2}C_u \\ &\quad\times C_1\Big\{\frac{C_b}{4}\sum_{q=0}^{p}\left(\frac{BR}{2}\right)^{p-q}K^{q-p-2} + K^{-1} + K^{-2}\Big\} \\ &\le C_1\varepsilon^{-2}M_{R,p}(f) + \frac{\max([p],R/\varepsilon)^{p+2}}{p!}K^{p+2}C_u \\ &\quad\times C_1\Big\{\frac{C_b}{4}\frac{1}{1-BR/(2K)}K^{-2} + K^{-1} + K^{-2}\Big\}. \end{aligned}$$

Finally, we have the bound

$$\begin{aligned} M_{R,p}(f) &\le \frac{1}{p!}C_f\gamma^p\left(\frac{R}{2}\right)^{2+p}\left[p^pR^{-p} + \max(p,\varepsilon^{-1})^p\right] \\ &\le \left(\frac{\gamma}{2}\right)^p\frac{1}{p!}R^2C_f\frac{1}{4}\left[p^p + \max(Rp,R/\varepsilon)^p\right]. \end{aligned}$$

As $R \le 1$, we get $p^p + \max(Rp,R/\varepsilon)^p \le 2\max(p,R/\varepsilon)^p$ and, together with $(R/\varepsilon)^2 \le \min(1,(R/\varepsilon)^2)\max([p],(R/\varepsilon)^2)$, we can conclude

$$C_1\varepsilon^{-2}M_{R,p}(f) \le C_1\frac{1}{2}\left(\frac{\gamma}{2}\right)^p C_f\min(1,(R/\varepsilon)^2)\,\frac{\max([p],R/\varepsilon)^{p+2}}{p!}, \tag{3.17}$$

$$\begin{aligned} N_{R,p}(u) &\le K^{p+2}\frac{\max([p],R/\varepsilon)^{p+2}}{p!}C_u \\ &\quad\times\left[C_1\frac{1}{2}\left(\frac{\gamma}{2K}\right)^pK^{-2} + C_1\Big\{\frac{C_b/4}{1-BR/(2K)}K^{-2} + K^{-1} + K^{-2}\Big\}\right]. \end{aligned}$$

The fact that the bracketed expression is bounded by one by the choice of K in (3.16) concludes the induction argument. ∎

3.2.2. Estimates at the boundary (proof of Proposition 3.4). The strategy for proving Proposition 3.4 is first to get control over the tangential derivatives of the solution u of (3.11), i.e., the x-derivatives of u. This will be accomplished in Lemma 3.10. In the second step, the remaining normal derivatives, i.e., the y-derivatives, will be controlled.

In order to carry out this two-step approach, we introduce the following notation for smooth functions u:

$$\begin{aligned} M'_{R,p}(u) &= \frac{1}{p!}\sup_{R/2\le r<R}(R-r)^{p+2}\|\partial_x^p u\|_{L^2(G_r)}, \\ N'_{R,p}(u) &= \begin{cases} \dfrac{1}{p!}\displaystyle\sup_{R/2\le r<R}(R-r)^{p+2}\|\nabla^2\partial_x^p u\|_{L^2(G_r)} & \text{if } p\ge 0, \\ \displaystyle\sup_{R/2\le r<R}(R-r)^{p+2}\|\nabla^{2+p}u\|_{L^2(G_r)} & \text{if } p=-2,-1, \end{cases} \\ \tilde{M}_{R,p}(u) &= \frac{1}{p!}\sup_{R/2\le r<R}(R-r)^{p+2}\|\nabla^p u\|_{L^2(G_r)}. \end{aligned}$$

Note that we have $N'_{R,p,0} \le N'_{R,p}$. Lemma 3.9 is the analogue of Lemma 3.7.

LEMMA 3.9. *Let $u \in H^1(G_R)$ solve $\Delta u = f$ on G_R and assume that $u = 0$ on $\partial G_R \cap \{(x,y)\,|\,y=0\}$. Then there is a generic constant $C_2 > 0$ such that*

$$N'_{R,p}(u) \le C_2 \left\{ M'_{R,p}(f) + N'_{R,p-1}(u) + N'_{R,p-2}(u) \right\}.$$

Proof. The proof can be found in [8, Lem. 5.7.3′]. ∎

We start with a bound on the tangential derivatives.

LEMMA 3.10. *Assume the hypotheses of Proposition* 3.4. *Then there is $K_1 > 0$ independent of ε and R such that with C_u of* (3.15)

$$N'_{R,p}(u) \le C_u K_1^{p+2} \frac{\max([p], R/\varepsilon)^{p+2}}{[p]!}, \qquad p \ge -2.$$

Proof. The proof is almost verbatim the same as the proof of Proposition 3.3. Instead of using Lemma 3.7 we make use of Lemma 3.9. In particular, the constant K_1 will be chosen such that $K_1 > \max(1, \gamma/2, BR/2)$. ∎

Proof of Proposition 3.4. Let K_1 be the constant of Lemma 3.10 and choose $K_2 > \max(1, BR/2, \gamma/2)$ such that $\forall\ p \ge 0,\ q \ge 0$

$$\begin{aligned} &\Big[\frac{1}{2}\left(\frac{\gamma}{2K_1}\right)^p \left(\frac{\gamma}{2K_2}\right)^q K_1^{-2} K_2^{-2} \\ &\quad + K_1^2 K_2^{-2} + \frac{C_b/4}{(1 - BR/(2K_1))(1 - BR/(2K_2))} K_2^{-2}\Big] \le 1. \end{aligned} \tag{3.18}$$

We will proceed by induction on q. By Lemma 3.10 and our earlier observation that $N'_{R,p,0} \le N'_{R,p}$, the claim (3.14) is true for $q = 0$ and all $p \ge 0$, and it is easy to see that the claim is also true for $q = -2$, $q = -1$: We have for $q = -2$ and $q = -1$ and $p \ge 0$ by a straightforward calculation that

$$\begin{aligned} N'_{R,p,-2}(u) &= \frac{1}{[p-2]!} \sup_{R/2 \le r < R} (R-r)^p \|\partial_x^p u\|_{L^2(G_r)} \le N'_{R,p-2}(u), \\ N'_{R,p,-1}(u) &= \frac{1}{[p-1]!} \sup_{R/2 \le r < R} (R-r)^{p+1} \|\partial_y\, \partial_x^p u\|_{L^2(G_r)} \le N'_{R,p-1}(u). \end{aligned}$$

Let us now proceed with the induction argument on q and assume that the induction hypothesis is proven for $-2 \le q' < q$. We have

$$\begin{aligned} -\partial_y^2 u &= \partial_x^2 u + \varepsilon^{-2}(f - bu), \\ |\partial_x^p\, \partial_y^{q+2} u| &\le |\partial_x^{p+2}\, \partial_y^q u| + \varepsilon^{-2} |\partial_x^p\, \partial_y^q f| + \varepsilon^{-2} |\partial_x^p\, \partial_y^q (bu)|. \end{aligned}$$

Leibniz's formula and the assumptions on b yield

$$\begin{aligned} |\partial_x^p\, \partial_y^q (bu)| &\le \sum_{m=0}^{q} \sum_{n=0}^{p} \binom{p}{n} \binom{q}{m} |\partial_x^{p-n}\, \partial_y^{q-m} b|\, |\partial_x^n\, \partial_y^m u| \\ &\le C_b \sum_{m=0}^{q} \sum_{n=0}^{p} \binom{p}{n} \binom{q}{m} B^{p+q-m-n} (p-n)!(q-m)! |\partial_x^n\, \partial_y^m u| \\ &\le C_b \sum_{m=0}^{q} \sum_{n=0}^{p} B^{p+q-n-m} (p+q)^{p+q-n-m} |\partial_x^n\, \partial_y^m u|, \end{aligned}$$

where, in the last step, we used the bound

$$\binom{p}{n}\binom{q}{m}(p-n)!(q-m)! = \frac{p!}{n!}\frac{q!}{m!} \le p^{p-n}q^{q-m} \le (p+q)^{p+q-n-m}.$$

Hence, we obtain for $N'_{R,p,q}(u)$

$$\begin{aligned}
&N'_{R,p,q}(u) = \frac{1}{[p+q]!}\sup_{R/2\le r<R}(R-r)^{p+q+2}\|\partial_x^p\,\partial_y^{q+2}\,u\|_{L^2(G_r)}\\
&\le N'_{R,p+2,q-2}(u) + \varepsilon^{-2}\tilde{M}_{R,p+q}(f)\\
&+ C_b\varepsilon^{-2}\sum_{m=0}^{q}\sum_{n=0}^{p}\left(\frac{BR}{2}\right)^{p+q-m-n}\left(\frac{R}{2}\right)^2\frac{(p+q)^{p+q-n-m}![m-2+n]!}{[p+q]!}N'_{R,n,m-2}(u).
\end{aligned}$$

By the induction hypothesis we have with C_u of (3.15)

$$\begin{aligned}
[m-2+n]!N_{R,n,m-2}(u) &\le C_u K_1^{n+2}K_2^m\max\left([m-2+n],R/\varepsilon\right)^{m+n}\\
&\le C_u K_1^{n+2}K_2^m\max\left([p+q],R/\varepsilon\right)^{m+n}
\end{aligned}$$

and with the bound $(R/\varepsilon)^2 \le \max\left([p+q],R/\varepsilon\right)^2$ we obtain

$$\begin{aligned}
N'_{R,p,q}(u) &\le K_1^{p+4}K_2^q\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{[p+q]!}C_u + \varepsilon^{-2}\tilde{M}_{R,p+q}(f)\\
&\quad+\frac{C_b}{4}C_u\sum_{m=0}^{q}\sum_{n=0}^{p}\left(\frac{BR}{2}\right)^{p+q-m-n}K_1^{n+2}K_2^m\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{(p+q)!}\\
&\le K_1^{p+4}K_2^q\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{[p+q]!}C_u + \varepsilon^{-2}\tilde{M}_{R,p+q}(f)\\
&\quad+C_uK_1^{p+2}K_2^q\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{[p+q]!}\frac{C_b}{4}\sum_{m=0}^{q}\sum_{n=0}^{p}\left(\frac{BR}{2}\right)^{p+q-m-n}K_1^{n-p}K_2^{m-q}\\
&\le \varepsilon^{-2}\tilde{M}_{R,p+q}(f) + K_1^{p+2}K_1^{q+2}\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{[p+q]!}C_u\\
&\quad\times\Big[K_1^2K_2^{-2}+\frac{C_b}{4}\frac{1}{(1-BR/(2K_1))(1-BR/(2K_2))}K_2^{-2}\Big].
\end{aligned}$$

Reasoning as in (3.17) we get

$$\varepsilon^{-2}\tilde{M}_{R,p+q}(f) \le \frac{1}{2}\left(\frac{\gamma}{2}\right)^{p+q}C_f\min\left(1,(R/\varepsilon)^2\right)\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{[p+q]!}$$

and may therefore conclude that

$$\begin{aligned}
&N'_{R,p,q}(u) \le K_1^{p+2}K_2^{q+2}\frac{\max\left([p+q],R/\varepsilon\right)^{p+q+2}}{[p+q]!}C_u\\
&\times\left[\frac{1}{2}\left(\frac{\gamma}{2K_1}\right)^p\left(\frac{\gamma}{2K_2}\right)^q K_1^{-2}K_2^{-2}+K_1^2K_2^{-2}+\frac{C_b/4\,K_2^{-2}}{(1-BR/(2K_1))(1-BR/(2K_2))}\right].
\end{aligned}$$

As the bracketed expression is bounded by one by the choice of K_2 in (3.18), the induction argument is completed. ∎

3.2.3. Proof of Lemma 3.6. The growth conditions on the derivatives of f imply that f can be extended to a holomorphic function (also denoted f) on $\tilde{G} \subset \mathbb{C} \times \mathbb{C}$ with $\overline{G} \subset \tilde{G}$ and $\tilde{G}$ independent of $\varepsilon > 0$. First, we claim that there are δ_0, γ', $C > 0$ depending only on γ and C_f such that

$$\|f(\cdot + z_1(\cdot), \cdot + z_2(\cdot))\|_{L^2(G)} \leq C e^{\gamma' \delta/\varepsilon} \tag{3.19}$$

for all continuous functions $z_1,\ z_2 : G \to \mathbb{C}$ with $\|z_i\|_{L^\infty(G)} \leq \delta \leq \delta_0,\ i = 1, 2$. As f is holomorphic on $\tilde{G}$, there is $\delta_0 > 0$ such that $\forall\ (x, y) \in \overline{G}$ the power series expansion of f about (x, y) converges on a ball of radius $2\delta_0$. For functions $z_1,\ z_2$ with $\|z_i\|_{L^\infty(G)} \leq \delta \leq \delta_0$ we obtain

$$|f(x + z_1(x, y), y + z_2(x, y))| = \left| \sum_{\alpha \in \mathbb{N}_0^2} \frac{1}{\alpha!} D^\alpha f(x, y)(z_1, z_2)^\alpha \right| \leq \sum_{\alpha \in \mathbb{N}_0^2} \frac{1}{\alpha!} \left| D^\alpha f(x, y) \right| \delta^{|\alpha|}.$$

Therefore we get

$$\begin{aligned}
&\|f(\cdot + z_1(\cdot), \cdot + z_2(\cdot))\|_{L^2(G)} \leq \sum_{\alpha \in \mathbb{N}_0^2} \frac{1}{\alpha!} \|D^\alpha f\|_{L^2(G)} \delta^{|\alpha|} \\
&\leq \sum_{p=0}^{\infty} \sum_{|\alpha|=p} \left((p!)^{1/2} (\alpha!)^{-1/2} \|D^\alpha f\|_{L^2(G)} \right) \left((\alpha!)^{-1/2} p!^{-1/2} \delta^p \right) \\
&\leq \sum_{p=0}^{\infty} \|\nabla^p f\|_{L^2(G)} \left(\sum_{|\alpha|=p} \frac{1}{\alpha! p!} \delta^{2p} \right)^{1/2} = \sum_{p=0}^{\infty} \|\nabla^p f\|_{L^2(G)} \frac{1}{p!} 2^{p/2} \delta^p \\
&\leq C_f \sum_{0 \leq p \leq \varepsilon^{-1}} \frac{1}{p!} \left(\sqrt{2} \gamma \varepsilon^{-1} \delta \right)^p + C_f \sum_{p > \varepsilon^{-1}} \frac{p^p}{p!} \gamma^p 2^{p/2} \delta^p \\
&\leq C_f e^{\sqrt{2}\gamma\delta/\varepsilon} + C \sum_{p > \varepsilon^{-1}} \left(e \sqrt{2} \gamma \delta \right)^p \leq C_f e^{\sqrt{2}\gamma\delta/\varepsilon} + \frac{1}{1 - \sqrt{2}\gamma\delta_0} \leq C e^{\sqrt{2}\gamma\delta/\varepsilon},
\end{aligned}$$

where we used Stirling's formula in the form $p! \geq C p^p e^{-p}$ and made the tacit assumption that δ_0 is so small that $e\sqrt{2}\gamma\delta_0 < 1$ so that the second sum is finite. This proves (3.19).

As g is analytic on $\overline{G}_1$ there is a holomorphic extension (also denoted g) to $\tilde{G}_1 \subset \mathbb{C} \times \mathbb{C}$. Thus, there are $\eta,\ \delta_0' > 0$ such that $\forall (x, y) \in \overline{G}_1$

$$|g_i(x + z_1, y + z_2) - g_i(x, y)| \leq \eta\delta, \qquad i = 1, 2, \quad z_1, z_2 \in \mathbb{C} \text{ with } |z_1|, |z_2| \leq \delta \leq \delta_0'. \tag{3.20}$$

For any $0 < \delta \leq \min(\delta_0', \delta_0/\eta)$ we obtain by Cauchy's integral theorem for derivatives for every $(x, y) \in G_1$ and every $\alpha = (\alpha_1, \alpha_2) \in \mathbb{N}_0^2$

$$\begin{aligned}
D^\alpha (f \circ g)(x, y) &= -\frac{\alpha!}{4\pi^2} \int_{|z_1|=\delta} \int_{|z_2|=\delta} \frac{(f \circ g)(x + z_1, y + z_2)}{z_1^{\alpha_1+1} z_2^{\alpha_2+1}} \, dz_1 dz_2, \\
|D^\alpha (f \circ g)(x, y)|^2 &\leq \frac{\alpha!^2}{4\pi^2 \delta^{2|\alpha|+2}} \int_{|z_1|=\delta} \int_{|z_2|=\delta} |(f \circ g)(x + z_1, y + z_2)|^2 \, |dz_1| \, |dz_2|.
\end{aligned}$$

By (3.20), we can write

$$g_1(x+z_1,y+z_2)=g_1(x,y)+\zeta_1,\qquad g_2(x+z_1,y+z_2)=g_2(x,y)+\zeta_2,$$

where ζ_1, ζ_2 are smooth functions of x, y, z_1, z_2, and $|\zeta_i|\le\eta\delta$, $i=1,2$. Integrating over G_1, we obtain after the smooth change of variables $g(x,y)=(x',y')$ (note that $0<c_1\le|\det g'|\le c_2<\infty$) and denoting ζ_1', ζ_2' the functions corresponding to ζ_1, ζ_2 after the change of variables

$$\begin{aligned}&|D^\alpha(f\circ g)(x,y)|^2_{L^2(G_1)}\\&\le c_2\frac{(\alpha!)^2}{4\pi^2\delta^{2|\alpha|+2}}\int_{|z_1|=\delta}\int_{|z_2|=\delta}\int_G|f(x'+\zeta_1',y'+\zeta_2')|^2\,dx'dy'|dz_1|\,|dz_2|.\end{aligned}$$

As $|\zeta_1'|$, $|\zeta_2'|\le\eta\delta$ uniformly in $(x',y')\in G$, $|z_1|$, $|z_2|\le\delta$, estimate (3.19) yields

$$\|D^\alpha(f\circ g)\|_{L^2(G_1)}\le C\frac{\alpha!}{\delta^{|\alpha|}}e^{\gamma'\eta\delta/\varepsilon}\qquad\forall 0<\delta\le\min(\delta_0',\delta_0/\eta).$$

In order to extract from this estimate the claim of the lemma, we distinguish the cases $|\alpha|\varepsilon$ large and $|\alpha|\varepsilon$ small. If $|\alpha|\varepsilon/(\eta\gamma')<\min(\delta_0',\delta_0/\eta)$, choose $\delta:=|\alpha|\varepsilon/(\eta\gamma')$ to get with Stirling's formula

$$\|D^\alpha(f\circ g)\|_{L^2(G_1)}\le C(\eta\gamma')^{|\alpha|}\sqrt{1+|\alpha|}\,\varepsilon^{-|\alpha|}.$$

If $|\alpha|\varepsilon/(\eta\gamma')\ge\min(\delta_0',\delta_0/\eta)$, choose $\delta:=\min(\delta_0',\delta_0/\eta)$ and observe that this implies $\delta\eta\gamma'\varepsilon^{-1}\le|\alpha|$ to arrive at

$$\|D^\alpha(f\circ g)\|_{L^2(G_1)}\le C\alpha!\;\delta^{-|\alpha|}e^{|\alpha|},$$

which completes the proof of Lemma 3.6. ∎

3.3. Proof of Theorem 3.1. Let $B_R(x_0)\subset\Omega$ be a ball of radius $R\le1$. Proposition 3.3 yields the existence of C, $K>0$ independent of ε, p such that for all $p\in\mathbb{N}_0$

$$\|\nabla^p u_\varepsilon\|_{L^2(B_{R/2}(x_0))}\le CK^p\max(p,\varepsilon^{-1})^p\left(1+\|u_\varepsilon\|_{L^2(B_R(x_0))}+\varepsilon\|\nabla u_\varepsilon\|_{L^2(B_R(x_0))}\right).$$

Let us now consider estimates at the boundary. First, we see that we may consider the case of homogeneous Dirichlet data: As the boundary data g is analytic, it can be extended analytically into Ω, e.g., by taking as the extension function the function G defined by

$$-\Delta G=0\quad\text{ on }\Omega,\qquad G=g\quad\text{ on }\partial\Omega.$$

As $\partial\Omega$ and g are assumed to be analytic, standard elliptic theory [8, 9] gives that G is analytic on a neighborhood of $\overline{\Omega}$. Note that G is independent of ε. The auxiliary function $\tilde u=u-G$ solves

$$\begin{aligned}-\varepsilon^2\Delta\tilde u+\tilde u=\tilde f:=f+\varepsilon^2\Delta G-G=f-G\qquad&\text{on }\Omega,\\\tilde u=0\qquad&\text{on }\partial\Omega\end{aligned}$$

and by the triangle inequality

$$\|\nabla^p u\|_{L^2(B\cap\Omega)}\le\|\nabla^p\tilde u\|_{L^2(B\cap\Omega)}+\|\nabla^p G\|_{L^2(B\cap\Omega)}\qquad\forall p\in\mathbb{N}_0$$

for balls B. It suffices therefore to get the desired bounds for $\tilde{u}$.

In order to apply Proposition 3.4, we introduce a mapping to flatten the boundary locally: For $R > 0$ and a point $x_0 \in \partial\Omega$, we introduce the conformal map ζ which maps $Q := \Omega \cap B_{2R}(x_0)$ conformally onto G_{2R}. The transformed functions $\hat{u} = \tilde{u} \circ \zeta^{-1}$, $\hat{f} = \tilde{f} \circ \zeta^{-1}$ then solve

$$\begin{aligned} -\varepsilon^2 \Delta \hat{u} + |(\zeta^{-1})'|^2 \hat{u} &= \hat{f} |(\zeta^{-1})'|^2 \qquad \text{on } G_{2R}, \\ \hat{u} &= 0 \qquad \text{on } \partial G_{2R} \cap \{(x,y) \,|\, y = 0\}. \end{aligned}$$

Furthermore, by the analyticity of $\partial\Omega$, the function $|(\zeta^{-1})'|^2$ is (real) analytic on G_{2R} and hence Proposition 3.4 is applicable (note that $\tilde{f}$ and hence $\hat{f}|(\zeta^{-1})'|^2$ are independent of ε), and we get the desired estimate for $\hat{u}$, i.e.,

$$\|\nabla^p \hat{u}\|_{L^2(G_{R/2})} \le C K^p \max(p, \varepsilon^{-1})^p \left(1 + \|\hat{u}\|_{L^2(G_R)} + \varepsilon \|\nabla \hat{u}\|_{L^2(G_R)}\right) \qquad \forall p \in \mathbb{N}_0.$$

Applying Lemma 3.6 allows us to infer a similar estimate for $\tilde{u}$:

$$\|\nabla^p \tilde{u}\|_{L^2(B \cap \Omega)} \le C K'^p \max(p, \varepsilon^{-1})^p \left(1 + \|u_\varepsilon\|_{L^2(Q)} + \varepsilon \|\nabla u_\varepsilon\|_{L^2(Q)}\right) \qquad \forall p \in \mathbb{N}_0,$$

where B is a ball or radius $R' > 0$ with center x_0 such that $B_{R'} \cap \Omega \subset \zeta^{-1}(G_{R/2})$. The constants C, $K' > 0$ depend again on R, f, g, and the point x_0 but are independent of ε.

A compactness argument allows us to conclude the proof of the theorem. ∎

REFERENCES

[1] J. M. Melenk and C. Schwab, *hp FEM for reaction-diffusion equations I: Robust exponential convergence*, SIAM J. Numer. Anal., 35 (1998), pp. 1520–1557.

[2] J. L. Lions, *Perturbations singulières dans les problèmes aux limites et en contrôle optimal*, Lecture Notes in Math. 323, Springer-Verlag, Berlin, New York, 1973.

[3] W. Eckhaus, *Asymptotic Analysis of Singular Perturbations*, North-Holland, Amsterdam, 1979.

[4] D. N. Arnold and R. S. Falk, *Asymptotic analysis of the boundary layer for the Reissner–Mindlin plate model*, SIAM J. Math. Anal, 27 (1996), pp. 486–514.

[5] S.-D. Shih and R. B. Kellogg, *Asymptotic analysis of a singular perturbation problem*, SIAM J. Math. Anal, 18 (1987), pp. 1467–1511.

[6] H. Han and R. B. Kellogg, *Differentiability properties of solutions of the equation* $-\varepsilon^2 \Delta u + ru = f(x,y)$ *in a square*, SIAM J. Math. Anal, 21 (1990), pp. 394–408.

[7] R. B. Kellogg, *Boundary layers and corner singularities for a self-adjoint problem*, in Boundary Value Problems and Integral Equations in Non-smooth Domains, M. Costabel, M. Dauge, and S. Nicaise, eds., Lecture Notes in Pure and Appl. Math. 167, Marcel Dekker, New York, 1995, pp. 121–149.

[8] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1966.

[9] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications,* Vol. III, Dunod, Paris, 1968.

[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products,* corrected and enlarged edition, Academic Press, New York, 1980.

[11] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 401–430

ON STABILITY OF CONSERVATION LAWS*

GUNILLA KREISS†, HEINZ-OTTO KREISS‡, AND JENS LORENZ§

Abstract. We consider the Cauchy problem for systems of PDEs of the general form

$$u_t = P_0 u + \varepsilon_1 P_1 u + \varepsilon_2 Q(u) + \sum_j D_j F_j(x,t), \quad u = u(x,t).$$

Here P_0 has constant coefficients. The terms $\varepsilon_1 P_1 u$ and $\varepsilon_2 Q(u)$ describe linear and nonlinear perturbations, respectively, and $\sum_j D_j F_j(x,t)$ is a forcing term, which decays to zero for $t \to \infty$. The perturbation terms are assumed to have conservation form. We call the system nonlinearly stable if the solution $u(x,t)$ with $u(x,0) = 0$ remains smooth for all $t \geq 0$ and the maximum norm of u tends to zero for $t \to \infty$, provided that $\varepsilon_1^2 + \varepsilon_2^2$ is sufficiently small. In the paper we give sufficient conditions for nonlinear stability.

If the unperturbed system $u_t = P_0 u$ is parabolic, then the Laplace transform technique is satisfactory to derive conditions for nonlinear stability. However, if $u_t = P_0 u$ is hyperbolic or coupled parabolic-hyperbolic, then the Laplace transform technique fails if the perturbation terms are first-order differential operators. In this case, we combine Laplace transformation for small wave vectors with an energy technique to control the large-wave–number projection of the solution.

Key words. nonlinear stability, conservation laws, symmetrizer, Laplace transform

AMS subject classifications. 35G209

PII. S0036141097322479

1. Introduction.

We consider systems of partial differential equations of the form

$$u_t = P_0 u + \varepsilon_1 P_1 u + \varepsilon_2 \sum_{j=1}^{d} D_j g_j(x,t,u) + \sum_{j=1}^{d} D_j F_j(x,t), \quad x \in \mathbb{R}^d,\ t \geq 0, \tag{1.1}$$

with initial condition

$$u(x,0) = 0, \quad x \in \mathbb{R}^d. \tag{1.2}$$

Here $u = u(x,t)$ takes values in $\mathbb{R}^n$. The operator P_0 has constant coefficients,

$$P_0 = \sum_{|\alpha| \leq m} A_\alpha D^\alpha, \quad A_\alpha \in \mathbb{R}^{n \times n},$$
$$D^\alpha = D_1^\alpha \dots D_d^\alpha, \quad D_j = \partial/\partial x_j, \quad |\alpha| = \alpha_1 + \dots + \alpha_d.$$

The operator P_1 has conservation form,

$$P_1 u = \sum_{j=1}^{d} D_j \Big(B_j(x,t) u \Big)$$

*Received by the editors June 9, 1997; accepted for publication November 19, 1997; published electronically January 27, 1999.

http://www.siam.org/journals/sima/30-2/32247.html

†Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden (gunillak@nada.kth.se).

‡Department of Mathematics, UCLA, Los Angeles, CA 90095 (kreiss@math.ucla.ed). This author was supported by Office of Naval Research grant n00014 90 j 1382.

§Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131 (lorenz@math.unm.edu). This author was supported by NSF grant DMS-9404124 and DOE grant DE-FG03-95ER25235.

with variable coefficients

$$B_j(x,t) \in \mathbb{R}^{n\times n}.$$

The nonlinear functions $g_j(x,t,u)$ vanish quadratically at $u = 0$. More specific assumptions on $P_0,\ P_1,\ g_j(x,t,u)$, and $F_j(x,t)$ will be given below. Note that all terms in (1.1) have conservation form, except for a zero-order term in P_0, which is allowed.

We call (1.1), (1.2) nonlinearly stable if the solution $u(x,t)$ remains smooth for all $t \geq 0$ and the maximum norm $|u(\cdot,t)|_\infty$ tends to zero for $t \to \infty$, provided that $\varepsilon_1^2 + \varepsilon_2^2$ is sufficiently small. The aim of the paper is to give sufficient conditions on P_0, P_1, g_j, F_j, which imply nonlinear stability.

Our results can easily be translated into nonlinear stability results for problems with more general initial data than (1.2); see section 4.1.

We now outline the results of our paper. In section 2 we will first assume that the constant coefficient operator P_0 has the form

$$P_0 = \Delta + \sum_{j=1}^{d} A_j D_j, \quad A_j \in \mathbb{R}^{n\times n},$$

where $\Delta = D_1^2 + \cdots + D_d^2$ is the Laplacian and where the corresponding system

$$u_t = \sum_j A_j D_j u$$

is strongly hyperbolic. The linear problem

$$u_t = P_0 u + \sum_j D_j F_j(x,t), \quad u = 0 \text{ at } t = 0, \tag{1.3}$$

can be discussed by Fourier–Laplace transformation. If

$$\tilde{u}(\omega,s) = (2\pi)^{-d/2} \int_0^\infty \int_{\mathbb{R}^d} e^{-st} e^{-i\omega\cdot x} u(x,t)dx\,dt$$

denotes the Fourier–Laplace transform, then (1.3) becomes

$$s\tilde{u} = \hat{P}_0(\omega)\tilde{u} + i\sum_j \omega_j \tilde{F}. \tag{1.4}$$

Throughout the paper, the derivation of solution estimates for small $|\omega|$ (long wavelengths) and large $|\omega|$ (short wavelengths) will proceed quite differently. We therefore decompose

$$\tilde{u} = \tilde{u}^I + \tilde{u}^{II}$$

with

$$\tilde{u}^I(\omega,s) = \begin{cases} \tilde{u}(\omega,s) & \text{if } |\omega| \leq 1, \\ 0 & \text{if } |\omega| > 1, \end{cases} \tag{1.5}$$

and obtain a corresponding decomposition of u,

$$u(x,t) = u^I(x,t) + u^{II}(x,t),$$

by inverting the Fourier–Laplace transform. (Note that the time dependence is irrelevant in determining the decomposition $u = u^I + u^{II}$. The decomposition is well defined for any L_2-function $u = u(x)$.) For operators

$$P_0 = \Delta + \sum_j A_j D_j, \tag{1.6}$$

solution estimates for the linear problem (1.3) can then be derived from the transformed equation (1.4). The estimate is stated in Theorem 2.1. To discuss the nonlinear problem (1.1), it suffices to apply the linear estimate with F_j replaced by

$$F_j + \varepsilon_1 B_j u + \varepsilon_2 g_j(x,t,u).$$

Then, using nothing more than Sobolev inequalities, we obtain nonlinear stability. The details are given in Theorem 2.2. At the end of section 2 we briefly discuss how to extend the results to more general parabolic systems (1.1).

In section 3 we drop the assumption that $u_t = P_0 u$ is parabolic and consider a general constant coefficient operator $P_0 = \sum_{|\alpha|\le m} A_\alpha D^\alpha$. We then formulate our assumptions on P_0 in terms of its symbol,

$$\hat{P}_0(\omega) = \sum_{|\alpha|\le m} A_\alpha (i\omega_1)^{\alpha_1} \cdots (i\omega_d)^{\alpha_d}, \quad \omega \in \mathbb{R}^d.$$

A main assumption will be the following eigenvalue condition.

Assumption 1. There is a constant $c_0 > 0$ such that

$$\operatorname{Re}\lambda \le \begin{cases} -c_0|\omega|^2 & \text{if } |\omega| \le 1, \\ -c_0 & \text{if } |\omega| \ge 1 \end{cases}$$

for all eigenvalues λ of $\hat{P}_0(\omega)$.

For the other assumptions, see section 3. Clearly, for the operator $P_0 = \Delta + \sum_j A_j D_j$ discussed in section 2, one has

$$\operatorname{Re}\lambda = -|\omega|^2 \quad \text{for all} \quad \lambda \in \sigma\Big(\hat{P}_0(\omega)\Big)$$

without restricting $|\omega|$. In Assumption 1 such a behavior of $\operatorname{Re}\lambda$ is required for small $|\omega|$ but not for large $|\omega|$. Therefore, the assumptions on P_0 in section 3 allow for systems $u_t = P_0 u$, which are coupled parabolic-hyperbolic or strongly hyperbolic with suitable zero-order terms. (In the appendix we give sufficient conditions which imply all the requirements on P_0 of section 3.)

Since the upper bound

$$\operatorname{Re}\lambda \le -c_0 < 0 \quad \text{for all} \quad \lambda \in \sigma\Big(\hat{P}_0(\omega)\Big), \quad |\omega| \ge 1,$$

on $\operatorname{Re}\lambda$ in Assumption 1 does not tend to $-\infty$ for $|\omega| \to \infty$, the perturbation terms

$$\varepsilon_1 \sum_j D_j(B_j u) + \varepsilon_2 \sum_j D_j g_j(x,t,u) \tag{1.7}$$

cannot be treated as forcing terms anymore for large $|\omega|$, because they are unbounded operators in L_2. Instead, to derive estimates for the large-$|\omega|$ part u^{II} of the solution, we will employ an energy estimate and consider

$$\frac{d}{dt} \|u^{II}(\cdot,t)\|^2_{\mathcal{H}} .$$

Here the $\mathcal{H}$-norm will be determined by an inner product

$$(u,v)_{\mathcal{H}} = \int_{\mathbb{R}^d} \hat{u}^*(\omega) H(\omega) \hat{v}(\omega) d\omega,$$

where

$$\hat{u}(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\omega\cdot x} u(x) dx$$

is the Fourier transform and where $H(\omega)$ is a symmetrizer for $\hat{P}_0(\omega)$. The main idea, then, of section 3 is to combine the Laplace transform technique for $|\omega| \leq 1$ with energy estimates for $|\omega| > 1$.

Remark. One can try, of course, to use an energy technique for the small–$|\omega|$ part of the solution as well. However, since generally $\mathrm{Re}\lambda \approx -c_0|\omega|^2$ for $\lambda \in \sigma(\hat{P}(\omega))$ if $|\omega|$ is small, there is no exponential decay in time which is uniform in $|\omega|$. This is a well-known difficulty, and there are approaches different from ours to deal with it. See, for example, [2] and [5].

In section 3, we assume that the symmetrizer $H(\omega)$ will only depend on the symbol $\hat{P}_0(\omega)$ of the constant coefficient operator P_0 and will tend to the identity for $|\omega| \to \infty$,

$$|H(\omega) - I| \leq \frac{\text{const.}}{|\omega|}, \quad |\omega| \geq 1. \tag{1.8}$$

(For such a condition, see [1].) Then, to treat the perturbation terms (1.7), we will assume that the matrices $B_j(x,t)$ and $g_{ju}(x,t,u)$ are symmetric.

Symmetrizers and energy estimates can be used more generally; see, for example, [4]. Also, the combination of the Laplace transform technique for $|\omega| \leq 1$ with an energy technique for $|\omega| > 1$ does not depend on the specific properties of $H(\omega)$ in section 3. Therefore, condition (1.8) as well as the symmetry assumptions for B_j and g_{ju} can be modified. We formulate a corresponding result in section 4 as a conjecture. Details will be provided in a forthcoming paper.

Notations.

1.
$$|u|^2 = \sum_j |u_j|^2 : \quad \text{Euclidean norm;}$$

2.
$$|A| = \max\{|Au| : |u| = 1\} : \quad \text{corresponding matrix norm;}$$

3.
$$\hat{u}(\omega) = (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{-i\omega\cdot x} u(x) dx : \quad \text{Fourier transform;}$$

4.
$$u = u^I + u^{II} \text{ where } \hat{u}^I(\omega) = \begin{cases} \hat{u}(\omega) & \text{if } |\omega| \leq 1, \\ 0 & \text{if } |\omega| > 1 \end{cases} :$$

decomposition of u into a small-wave–vector projection u^I and a large-wave–vector projection u^{II};

5.

$$\tilde{u}(\omega, s) = (2\pi)^{-d/2} \int_0^\infty \int_{\mathbb{R}^d} e^{-st} e^{-i\omega\cdot x} u(x,t)dx\,dt : \quad \text{Fourier–Laplace transform;}$$

6.

$$(u, v) = \int_{\mathbb{R}^d} u^*(x)v(x)dx = \int_{\mathbb{R}^d} \hat{u}^*(\omega)\hat{v}(\omega)d\omega :$$

L_2-inner product of $u, v \in L_2(\mathbb{R}^d, \mathbb{R}^n)$;

7.

$$(u, v)_{\mathcal{H}} = \int_{\mathbb{R}^d} \hat{u}^*(\omega)H(\omega)\hat{v}(\omega)d\omega :$$

modified inner product determined by symmetrizer $H(\omega)$;

8.

$$\|u\|^2 = (u, u), \quad \|u\|^2_{\mathcal{H}} = (u, u)_{\mathcal{H}} : \quad \text{corresponding norms;}$$

9.

$$\|u\|^2_{H^p} = \sum_{|\alpha|\le p} \|D^\alpha u\|^2, \quad \|u\|^2_{p,\mathcal{H}} = \sum_{|\alpha|\le p} \|D^\alpha u\|^2_{\mathcal{H}} :$$

Sobolev and modified Sobolev norm;

10.

$$M(F, T) = \sum_j \left(\int_0^T \int_{\mathbb{R}^d} |F_j(x,t)|dxdt \right)^2 : \quad \text{square of } L_1\text{-norm;}$$

11.

$$|||\ B|||^{\ 2} = \int_0^\infty \int_{\mathbb{R}^d} |B(x,t)|^2 dxdt \ : \quad L_2\text{-norm over space-time;}$$

12.

$$\sigma(A) = \text{set of eigenvalues of } A;$$

13.

$$\mathrm{Re}A = \frac{1}{2}(A + A^*) : \quad \text{symmetric part of matrix } A;$$

14.

$$|\cdot|_\infty : \quad \text{sup-norm.}$$

2. The parabolic case. In this section we first assume that P_0 has the form

$$P_0 = \Delta + \sum_{j=1}^d A_j D_j, \quad A_j \in \mathbb{R}^{n\times n},$$

where the system $u_t = \sum_j A_j D_j u$ is strongly hyperbolic. That is, we make the following assumption.

Assumption 2. There is a constant $C > 0$ and, for each $\omega \in \mathbb{R}^d$ with $|\omega| = 1$, there is a transformation $S(\omega) \in \mathbb{C}^{n\times n}$ such that

a) $|S(\omega)| + |S^{-1}(\omega)| \le C$;
b) $S^{-1}(\omega)(\sum_j \omega_j A_j)S(\omega)$ is real, diagonal.

At the end of the section, we consider more general operators P_0 for which $u_t = P_0 u$ is parabolic. Next, we list our assumptions on the coefficients

$$B_j(x,t),\ g_j(x,t,u),\ F_j(x,t), \quad j = 1, \dots, d.$$

Assumption 3.
a) $B_j(x,t)$, $g_j(x,t,u)$, $F_j(x,t)$ are C^∞-functions defined for $x \in \mathbb{R}^d$, $t \ge 0$, $u \in \mathbb{R}^n$;
b) $F_j(x,0) \equiv 0$;
c) $\int_0^\infty \int_{\mathbb{R}^d} |F_j(x,t)| dx\, dt < \infty$;
d) for all $p = 0, 1, \dots$, we have that

$$\int_0^\infty \Big\{ \|F(\cdot,t)\|^2_{H^p} + \|F_t(\cdot,t)\|^2_{H^p} \Big\} dt < \infty;$$

e) $\int_0^\infty \|B_j(\cdot,t)\|^2 dt < \infty$;
f) for all α, there is C_α with

$$|D^\alpha B_j(x,t)| + |D^\alpha B_{jt}(x,t)| \le C_\alpha \quad \text{for} \quad x \in \mathbb{R}^d,\ t \ge 0;$$

g) for all α, β and $L > 0$, there is $C(\alpha,\beta,L)$ with

$$|D_x^\alpha D_u^\beta g(x,t,u)| + |D_x^\alpha D_u^\beta g_t(x,t,u)| \le C(\alpha,\beta,L) \quad \text{for} \quad x \in \mathbb{R}^d,\ t \ge 0,\ |u| \le L;$$

h) for all $L > 0$, there is C_L with

$$|g(x,t,u)| \le C_L |u|^2 \quad \text{for } x \in \mathbb{R}^d,\ t \ge 0,\ |u| \le L;$$

i) for all α and all $L > 0$, there is $C(\alpha, L)$ with

$$|D_x^\alpha g_j(x,t,u)| + |D_x^\alpha g_{jt}(x,t,u)| + |D_x^\alpha g_{ju}(x,t,u)| \le C(\alpha,L)|u| \quad \text{for} \quad x \in \mathbb{R}^d,\ t \ge 0,\ |u| \le L.$$

Remark. These assumptions might seem very restrictive, but they can often be realized by simple transformations; see the discussion in section 4. The assumption $u(x,0) = F_j(x,0) \equiv 0$ is convenient since we will use the Laplace transformation.

We first consider the problem (1.1), (1.2) with $\varepsilon_1 = \varepsilon_2 = 0$ and show the following estimate of u in terms of F.

THEOREM 2.1. *Consider the equation* $u_t = P_0 u + \sum_j D_j F_j(x,t)$, $u = 0$ *at* $t = 0$, *under Assumptions* 2–3. *Then, for any* $p = 1, 2, \dots$ *there is a constant* R_p, *independent of* T *and* F, *such that*

(2.1)
$$\int_0^T \Big\{ \|u\|^2_{H^p} + \|u_t\|^2_{H^p} \Big\} dt \le R_p \left\{ M(F,T) + \int_0^T \Big\{ \|F\|^2_{H^{p-1}} + \|F_t\|^2_{H^{p-1}} \Big\} dt \right\}.$$

Here

$$M(F,T) = \sum_j \Big(\int_0^T \int_{\mathbb{R}^d} |F_j(x,t)| dx\, dt \Big)^2.$$

Proof.
1) Fourier–Laplace transformation yields

$$s\tilde{u} = \hat{P}_0(\omega)\tilde{u} + i\sum_j \omega_j \tilde{F}_j, \quad \omega \in \mathbb{R}^d,\ s = \eta + i\xi. \tag{2.2}$$

We decompose

$$\tilde{u} = \tilde{u}^I + \tilde{u}^{II},$$

where $\tilde{u}^I(\omega, s) = \tilde{u}(\omega, s)$ for $|\omega| \le 1$ and $\tilde{u}^I(\omega, s) = 0$ for $|\omega| > 1$. Clearly,

$$\hat{P}_0 = -|\omega|^2 I + i\sum_j \omega_j A_j.$$

For $\omega \neq 0$, let $\omega^0 = \omega/|\omega|$ and use the transformation $S = S(\omega^0)$ of Assumption 2 to obtain

$$S^{-1}(sI - \hat{P}_0)S = (s + |\omega|^2)I - i\Lambda.$$

Here $\Lambda = \operatorname{diag}(\lambda_k)$ contains the real eigenvalues of $\sum_j \omega_j A_j$ as diagonal entries. From (2.2) we find that

$$\begin{aligned} |\tilde{u}|^2 &\le |(sI - \hat{P}_0)^{-1}|^2 |\omega|^2 |\tilde{F}|^2 \\ &\le C_1 |\omega|^2 |\tilde{F}|^2 \sum_{k=1}^{n} \frac{1}{(\eta + |\omega|^2)^2 + (\xi - \lambda_k)^2} \end{aligned} \tag{2.3}$$

with $s = \eta + i\xi,\ \eta \ge 0$. Here

$$|\tilde{F}(\omega, s)|^2 = \sum_j |\tilde{F}_j(\omega, s)|^2,$$

$$\tilde{F}_j(\omega, s) = (2\pi)^{-d/2} \int_0^\infty \int_{\mathbb{R}^d} e^{-st - i\omega\cdot x} F_j(x, t)\, dx\, dt,$$

and therefore,

$$|\tilde{F}(\omega, s)|^2 \le C_2\, M(F, \infty) \quad \text{for} \quad s = \eta + i\xi,\ \eta \ge 0.$$

For $|\omega| \le 1$ we obtain from (2.3), using Parseval's relation,

$$\begin{aligned} \int_0^\infty e^{-2\eta t} \|u^I\|^2 dt &= \frac{1}{2\pi} \int_{|\omega|\le 1} \int_{-\infty}^{\infty} |\tilde{u}(\omega, \eta + i\xi)|^2 d\xi\, d\omega \\ &\le C_3\, M(F, \infty) \sum_{k=1}^{n} I_k \end{aligned}$$

with

$$I_k = \int_{|\omega|\le 1} \int_{-\infty}^{\infty} \frac{|\omega|^2}{(\eta + |\omega|^2)^2 + (\xi - \lambda_k)^2}\, d\xi\, d\omega.$$

It is crucial that the integrals I_k are finite for $\eta \geq 0$. In fact, for $\eta = 0$,

$$I_k = \int_{|\omega|\leq 1}\int_{-\infty}^{\infty} \frac{|\omega|^2}{|\omega|^4+\xi^2}\, d\xi\, d\omega = \int_{|\omega|\leq 1}\int_{-\infty}^{\infty} \frac{d\xi'}{1+(\xi')^2}\, d\omega < \infty.$$

This shows that, for $\eta = 0$,

$$\int_0^\infty \|u^I\|^2 dt \leq C_4\, M(F,\infty).$$

To estimate space derivatives $D^\alpha u^I$, just note that

$$|\widetilde{(D^\alpha u^I)}(\omega,s)| \leq |\omega|^{|\alpha|}|u^I(\omega,s)| \leq |u^I(\omega,s)|.$$

Then Parseval's relation implies, as above,

$$\int_0^\infty \|u^I\|^2_{H^p} dt \leq C_5\, M(F,\infty), \quad C_5 = C_5(p).$$

Also, using (2.2), we obtain that

$$|\tilde{u}^I_t| = |s\tilde{u}^I| \leq C\Big\{|\tilde{u}^I| + |\tilde{F}|\Big\}, \quad |\omega| \leq 1,$$

and a similar estimate holds for $|\widetilde{(D^\alpha u^I_t)}|$. Therefore,

$$\int_0^\infty \|u^I_t\|^2_{H^p} dt \leq C_6\Big\{M(F,\infty) + \int_0^\infty \|F\|^2 dt\Big\}, \quad C_6 = C_6(p).$$

2) It remains to estimate u^{II}, i.e., to consider (2.2) for $|\omega| \geq 1$. Clearly, we obtain from (2.3), for $\eta = 0$,

$$(|\omega|^2+1)\,|\tilde{u}^{II}|^2 \leq C_7|\tilde{F}|^2, \quad |\omega| \geq 1.$$

Then Parseval's relation yields

$$\int_0^\infty \|u^{II}\|^2_{H^1} dt \leq C_8 \int_0^\infty \|F\|^2 dt.$$

Also, we can apply D^α and $D^\alpha \partial/\partial t$ to the given differential equation and obtain a corresponding estimate for $D^\alpha u^{II}$ and for $D^\alpha u^{II}_t$. Thus, our estimates show that (2.1) holds for $T = \infty$. Since values of $F(x,t)$ for $t > T$ do not affect the solution $u(x,t)$ for $t \leq T$, we can replace $T = \infty$ by for any finite T.

Remark. Global existence for the linear problem considered here is well known, and the solution can grow at most exponentially. Therefore, the formal process of Laplace transformation in t is justified for $s = \eta + i\xi$ if $\eta \geq \eta_0$, η_0 sufficiently large. Then, in inverting the Laplace transformation, our estimates show that no singularities are encountered for $\eta \geq 0$, and, therefore, the contour of integration can be deformed to $\eta = 0$. This justifies the formal use of the Laplace transform in t and the choice $\eta = 0$ in deriving solution estimates.

Now consider the nonlinear problem (1.1) with $P_0 = \Delta + \sum_j A_j D_j$ and recall Assumptions 2 and 3. For any choice of ε_1, ε_2, there is a local solution $u(x,t) = u(x,t,\varepsilon_1,\varepsilon_2)$, which is C^∞-smooth in a maximal interval

$$0 \leq t < \bar{T} = \bar{T}(\varepsilon_1,\varepsilon_2).$$

We will show that $\bar{T}(\varepsilon_1, \varepsilon_2) = \infty$ if $\varepsilon_1^2 + \varepsilon_2^2$ is sufficiently small. Furthermore, we will show that the solution tends to zero as $t \to \infty$.

THEOREM 2.2. *Consider* (1.1), (1.2) *where* $P_0 = \Delta + \sum_j A_j D_j$, *and recall Assumptions* 2 and 3. *There exists* $\varepsilon_0 > 0$ *such that the solution is* C^∞ *for all* $t \geq 0$ *if*

$$\varepsilon_1^2 + \varepsilon_2^2 \leq \varepsilon_0.$$

Furthermore,

$$\lim_{t\to\infty} |u(\cdot, t)|_\infty = 0.$$

Proof. Choose a large constant κ. (We will see below that the choice

$$\kappa = 1 + 2R_p \left\{ M(F, \infty) + \int_0^\infty \{\|F\|_{H^{p-1}}^2 + \|F_t\|_{H^{p-1}}^2\} dt \right\} \tag{2.4}$$

is sufficient. Here $p = d + 2$ and d is the number of space dimensions.)

Fix $\varepsilon_1,\ \varepsilon_2$ and let $u = u(x, t, \varepsilon_1, \varepsilon_2)$ denote the solution of (1.1), (1.2). Assume first that there is a finite time $T = T(\kappa, \varepsilon_1, \varepsilon_2)$ with

$$\int_0^T \{\|u\|_{H^p}^2 + \|u_t\|_{H^p}^2\} dt = \kappa. \tag{2.5}$$

Set

$$|D^\alpha u|_{\infty,T} = \sup_{x\in\mathbb{R}^d,\, 0\leq t\leq T} |D^\alpha u(x, t)|.$$

By Sobolev's inequality, there is a constant C, independent of T, such that

$$|D^\alpha u|_{\infty,T}^2 \leq C\kappa \quad \text{if} \quad |\alpha| + \frac{d}{2} < p. \tag{2.6}$$

(Here we have used that

$$\max_{0\leq t\leq T} |v(t)|^2 \leq C \int_0^T \{|v(t)|^2 + |v_t(t)|^2\} dt$$

with C independent of T if $v(0) = 0$.) Now apply Theorem 2.1 with F_j replaced by

$$F_j + \varepsilon_1 B_j u + \varepsilon_2 G_j, \quad G_j(x, t) = g_j(x, t, u(x, t)).$$

Then (2.5) and Theorem 2.1 yield

$$\begin{aligned}
\kappa &\leq R_p M(F + \varepsilon_1 Bu + \varepsilon_2 G, T) \\
&\quad + R_p \int_0^T \{\|F + \varepsilon_1 Bu + \varepsilon_2 G\|_{H^{p-1}}^2 + \|(F + \varepsilon_1 Bu + \varepsilon_2 G)_t\|_{H^{p-1}}^2\} dt \\
&\leq 2R_p \Big\{ M(F, T) + \int_0^T \{\|F\|_{H^{p-1}}^2 + \|F_t\|_{H^{p-1}}^2\} dt \Big\} \\
&\quad + 4\varepsilon_1^2 R_p M(Bu, T) \ + \ 4\varepsilon_2^2 R_p M(G, T) \int_0^T \{\|Bu\|_{H^{p-1}}^2 + \|(Bu)_t\|_{H^{p-1}}^2\} dt \\
&\quad + 4\varepsilon_2^2 R_p \int_0^T \{\|G\|_{H^{p-1}}^2 + \|G_t\|_{H^{p-1}}^2\} dt.
\end{aligned}$$

It remains to show that all terms multiplied by ε_1^2 or ε_2^2 can be controlled in terms of κ, if κ and T are related by (2.5). Then, by choosing κ as in (2.4) and making $\varepsilon_1^2 + \varepsilon_2^2$ small, we arrive at a contradiction to (2.4). This contradiction shows that a finite T with (2.5) cannot exist, and, therefore,

$$\int_0^T \left\{ \|u\|_{H^p}^2 + \|u_t\|_{H^p}^2 \right\} dt < \kappa \ , \quad T \geq 0.$$

Then standard arguments show that the solution exists and is C^∞ for all $t \geq 0$.

We now treat the terms multiplied by ε_j^2 on the right side of the above estimate of κ separately. We have

$$\left\{ \int_0^T \int_{\mathbb{R}^d} |B_j(x,t)| \, |u(x,t)| dx \, dt \right\}^2 \leq C_B \int_0^T \int_{\mathbb{R}^d} |u(x,t)|^2 dx \, dt \leq C_B \kappa,$$

using Assumption 3e. Furthermore, by Leibnitz' rule,

$$D^\alpha (B_j u) = \sum_{\beta \leq \alpha} c_{\alpha\beta} (D^{\alpha-\beta} B_j) D^\beta u,$$

where

$$\sup_{x,t} |D^{\alpha-\beta} B_j(x,t)| \leq \text{ const.}$$

by Assumption 3f. Therefore,

$$\int_0^T \|Bu\|_{H^{p-1}}^2 dt \leq C_B \kappa,$$

and similarly,

$$\int_0^T \|(Bu)_t\|_{H^{p-1}}^2 dt \leq C_B \kappa.$$

We now treat the nonlinear terms and recall

$$G_j(x,t) = g_j(x,t,u(x,t)).$$

First note that

$$\int_0^T \int_{\mathbb{R}^d} |G_j(x,t)| dx \, dt \leq C \int_0^T \int_{\mathbb{R}^d} |u(x,t)|^2 dx \, dt \leq C\kappa,$$

where we have used that g_j vanishes quadratically at $u = 0$. This implies $M(G,T) \leq C_1 \kappa^2$.

Now consider $\int_0^T \|G\|_{H^{p-1}}^2 dt$, and let α be a multi-index with $|\alpha| \leq p-1$. Then $D^\alpha g_j(x,t,u(x,t))$ is a sum of terms

$$\varphi_j(x,t,\alpha,\sigma) D^{\sigma_1} u \ \cdots D^{\sigma_r} u, \tag{2.7}$$

where $\sigma_1, \ldots, \sigma_r$ are multi-indices with

$$|\sigma_1| + \cdots + |\sigma_r| \leq |\alpha| \leq p-1;$$

the function φ_j is a partial derivative of g_j evaluated at $(x,t,u(x,t))$. The index r satisfies $0 \le r \le p-1$. Since $|u|^2_{\infty,T} \le C\kappa$ by (2.6), we have

$$\sup_{x,0\le t\le T} |\varphi_j| \le C_0(\kappa).$$

Consider a term (2.7) and first assume $1 \le r \le p-1$. By (2.6),

$$|D^{\sigma_j} u|_{\infty,T} \le C_0(\kappa) \quad \text{if} \quad |\sigma_j| + \frac{d}{2} < p,$$

and we say that $D^{\sigma_j}u$ is estimated in sup-norm in terms of κ. Suppose there are two factors in (2.7), $D^{\sigma_1}u$ and $D^{\sigma_2}u$, say, which *cannot* be estimated in sup-norm in terms of κ. Then we have

$$|\sigma_1| + \frac{d}{2} \ge p \quad \text{and} \quad |\sigma_2| + \frac{d}{2} \ge p;$$

thus,

$$p - 1 + d \ge |\sigma_1| + |\sigma_2| + d \ge 2p.$$

But this implies $p \le d-1$, which contradicts our choice of $p = d+2$. We conclude that each factor in (2.7), except at most one, can be estimated in sup-norm in terms of κ. Therefore,

$$\|\varphi_j D^{\sigma_1} u \ \cdots \ D^{\sigma_r} u\| \le C_0(\kappa)\|u\|_{H^{p-1}}.$$

If $r = 0$ in (2.7), then

$$\varphi_j = D^\alpha_x g_j(x,t,u),$$

and Assumption 3i implies

$$|\varphi_j| \le C|u|.$$

These arguments show that

$$\int_0^T \|G\|^2_{H^{p-1}} dt \le C_1(\kappa).$$

Now consider

$$\int_0^T \|G_t\|^2_{H^{p-1}} dt.$$

Since $G_j(x,t) = g_j(x,t,u(x,t))$ we have

$$G_{jt} = g_{jt} + g_{ju}u_t$$

and, therefore,

$$D^\alpha G_{jt} = D^\alpha g_{jt} + D^\alpha(g_{ju}u_t).$$

The term $D^\alpha g_{jt}$ is treated in the same way as $D^\alpha g_j$ above. Finally, $D^\alpha(g_{ju}u_t)$ is a sum of terms

$$\psi_j(x,t,\alpha,\sigma) D^{\sigma_1} u \ \cdots \ D^{\sigma_r} u D^\beta u_t \tag{2.8}$$

with $|\sigma_1|+\cdots+|\sigma_r|+|\beta| \le |\alpha| \le p-1$. The function ψ_j is a derivative of g_{ju} evaluated at $(x,t,u(x,t))$. First let

$$|\beta| + 2 + \frac{d}{2} < p. \tag{2.9}$$

Then we use the differential equation (1.1) to express $D^\beta u_t$ in terms of space derivatives of u of order $\le |\beta|+2$. Since (2.9) is assumed, all these space derivatives can be bounded in sup-norm in terms of κ. Therefore,

$$|D^\beta u_t|_{\infty,T} \le C(\kappa) \quad \text{if} \quad |\beta| + 2 + \frac{d}{2} < p.$$

If $|\beta| + 2 + \frac{d}{2} \ge p$ and $|\sigma_1| + \frac{d}{2} \ge p$, say, then

$$p + 1 + d \ge |\beta| + |\sigma_1| + 2 + d \ge 2p.$$

But this implies $p \le d+1$, which contradicts our choice $p = d+2$. Thus, if $|\beta|+2+\frac{d}{2} \ge p$, then $|\sigma_j| + \frac{d}{2} < p$, and, consequently, all terms $D^{\sigma_j}u$ in (2.8) are bounded in sup-norm in terms of κ. Also, if $r = 0$ in (2.8), then ψ_j is a space derivative $D_x^\gamma g_{ju}(x,t,u)$, and we use

$$|D_x^\gamma g_{ju}(x,t,u)| \le C|u|.$$

These arguments show that

$$\|\psi_j D^{\sigma_1}u \ \cdots \ D^{\sigma_r}uD^\beta u_t\|^2 \le C_2(\kappa)\big\{\|u\|_{H^p}^2 + \|u_t\|_{H^p}^2\big\},$$

and therefore,

$$\int_0^T \|G_t\|_{H^{p-1}}^2 dt \le C_3(\kappa).$$

To summarize, we define κ by (2.4) with $p = d+2$. If we assume that there is a finite $T = T(\kappa,\varepsilon_1,\varepsilon_2)$ with (2.5), then our linear estimate (Theorem 2.1) yields

$$\kappa \le C(\kappa)(\varepsilon_1^2+\varepsilon_2^2) + 2R_p\left\{M(F,T) + \int_0^T \big\{\|F\|_{H^{p-1}}^2 + \|F_t\|_{H^{p-1}}^2\big\}dt\right\}.$$

Choosing $\varepsilon_1^2+\varepsilon_2^2 \le \varepsilon_0$ with $\varepsilon_0 = 1/(2C(\kappa))$, we arrive at a contradiction to (2.4). Therefore, if $\varepsilon_1^2+\varepsilon_2^2 \le \varepsilon_0$, then

$$\int_0^\infty \big\{\|u\|_{H^p}^2 + \|u_t\|_{H^p}^2\big\}dt < \infty.$$

Consequently,

$$\max_{T\le t<\infty} |u(\cdot,t)|_\infty^2 \le C\int_T^\infty \big\{\|u\|_{H^p}^2 + \|u_t\|_{H^p}^2\big\}dt \to 0 \text{ as } T\to\infty,$$

and in particular,

$$\lim_{t\to\infty} |u(\cdot,t)|_\infty = 0.$$

This proves Theorem 2.2.

Remark. Throughout the paper, we do not try to minimize smoothness assumptions. In the definition of κ by (2.4) the choice $p = d + 2$ can be improved if one uses Hölder's inequality and a Gagliardo–Nirenberg inequality instead of Sobolev's inequality; see, for example, [1].

Generalization. It is not difficult to extend the results of this section to more general parabolic systems. Consider (1.1), (1.2) where P_0 is a constant coefficient operator

$$P_0 = \sum_{|\alpha| \le m} A_\alpha D^\alpha, \quad A_\alpha \in \mathbb{R}^{n \times n},$$

and assume that its symbol

$$\hat{P}_0(\omega) = \sum_{|\alpha| \le m} A_\alpha (i\omega_1)^{\alpha_1} \cdots (i\omega_d)^{\alpha_d}, \quad \omega \in \mathbb{R}^d,$$

satisfies the following two conditions.

1. There is a constant C_0 such that
$$|e^{\hat{P}_0(\omega)t}| \le C_0$$
for all $\omega \in \mathbb{R}^d$ and all $t \ge 0$.
2. There is a constant $c_0 > 0$ such that
$$\mathrm{Re}\lambda \le -c_0|\omega|^2$$
for all $\lambda \in \sigma(\hat{P}_0(\omega))$ and all $\omega \in \mathbb{R}^d$.

If, in addition, the terms B_j, g_j, F_j satisfy our general Assumption 3, then the result formulated in Theorem 2 is valid. To prove this, it suffices to show that the solution of the linear problem

$$u_t = P_0 u + \sum_j D_j F_j$$

satisfies the estimate (2.1). By the Kreiss' matrix theorem [3], there is a transformation $S = S(\omega)$ with

$$S^{-1}\hat{P}_0 S = \begin{pmatrix} \lambda_1 & r_{12} & \cdots & \cdots & r_{1n} \\ & \ddots & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & r_{n-1,n} \\ 0 & & & & \lambda_n \end{pmatrix}$$

$$|r_{jl}| \le C_1 |\mathrm{Re}\lambda_j|, \quad 1 \le j < l \le n, \quad |S| + |S^{-1}| \le C_1.$$

If $\lambda_j = \alpha_j + i\beta_j, \alpha_j, \beta_j \in \mathbb{R}$, then, by assumption, $\alpha_j \le -c_0|\omega|^2$, and one obtains

$$\left|\left((\eta + i\xi)I - \hat{P}_0\right)^{-1}\right|^2 \le C_2 \sum_j \frac{1}{(\eta + |\omega|^2)^2 + (\xi - \beta_j)^2}.$$

Therefore, we obtain again the estimate (2.3), and (2.1) follows. The remaining arguments are the same as in the proof of Theorem 2.

3. The symmetric case. In this section, let P_0 denote a general constant coefficient operator,

$$P_0 = \sum_{|\alpha| \le m} A_\alpha D^\alpha, \quad A_\alpha \in \mathbb{R}^{n \times n},$$

with symbol

$$\hat{P}_0(\omega) = \sum_{|\alpha| \le m} A_\alpha (i\omega_1)^{\alpha_1} \cdots (i\omega_d)^{\alpha_d}, \quad \omega \in \mathbb{R}^d.$$

We do not assume here that the system $u_t = P_0 u$ is parabolic and, therefore, we do not require the estimate

$$\mathrm{Re}\lambda \le -c_0 |\omega|^2, \quad \lambda \in \sigma(\hat{P}_0(\omega))$$

for large $|\omega|$. In this case, the technique of section 2 cannot be used to treat the perturbation

$$\varepsilon_1 \sum_j D_j (B_j u) + \varepsilon_2 \sum_j D_j g_j$$

for large $|\omega|$, because we do not "gain" a derivative in the estimate for the linear equation $u_t = P_0 u + \sum_j D_j F_j$. If one makes the assumptions:

1. There is a constant $c_0 > 0$ such that

$$\hat{P}_0(\omega) + \hat{P}_0^*(\omega) \le -2c_0 I < 0 \tag{3.1}$$

for all $\omega \in \mathbb{R}^d$;

2. $B_j(x,t) = B_j^T(x,t), \quad g_{ju}(x,t,u) = g_{ju}^T(x,t,u), \quad j = 1, \dots, d;$

then one can use standard energy estimates to derive the conclusions of Theorem 2. However, in many applications the strict negativity assumption (3.1) is not fulfilled. Consider, for example, the (simplified) compressible Navier–Stokes equations, linearized about a constant state U, V,

$$\begin{pmatrix} u \\ v \\ p \end{pmatrix}_t + \begin{pmatrix} U & 0 & 1 \\ 0 & U & 0 \\ 1 & 0 & U \end{pmatrix} \begin{pmatrix} u \\ v \\ p \end{pmatrix}_x + \begin{pmatrix} V & 0 & 0 \\ 0 & V & 1 \\ 0 & 1 & V \end{pmatrix} \begin{pmatrix} u \\ v \\ p \end{pmatrix}_y = \begin{pmatrix} \Delta u \\ \Delta v \\ 0 \end{pmatrix}. \tag{3.2}$$

A simple calculation shows that

$$\hat{P}_0(\omega) + \hat{P}_0^*(\omega) \le 0, \tag{3.3}$$

and there is a constant $c_0 > 0$ with

$$\mathrm{Re}\lambda \le -c_0 \frac{|\omega|^2}{1 + |\omega|^2}, \quad \lambda \in \hat{P}_0(\omega), \tag{3.4}$$

for all ω. If one assumes the above two conditions (3.3), (3.4) for $\hat{P}_0$, then one can estimate u^I as in section 2, since (3.3) implies $|e^{\hat{P}_0(\omega)t}| \le 1$ for all $t \ge 0$. To estimate

u^{II}, one can apply an energy estimate in a modified L_2-norm if additional conditions are satisfied.

This motivates the following requirement on P_0.

Assumption 4.

a) There is $C_0 > 0$ such that

$$|e^{\hat{P}_0(\omega)t}| \le C_0 \quad \text{for} \quad t \ge 0,\ |\omega| \le 1;$$

b) there is $c_0 > 0$ such that

$$\operatorname{Re}\lambda \le -c_0|\omega|^2 \quad \text{for all} \quad \lambda \in \sigma(\hat{P}_0(\omega)) \quad \text{if } |\omega| \le 1;$$

c) there are $c_1 > 0$ and $C_1 > 0$, and there is a smooth Hermitian matrix function $H(\omega)$ defined for $|\omega| \ge 1$ which satisfies the following three conditions:

1. $\frac{1}{C_1} I \le H(\omega) \le C_1 I$;
2. $\operatorname{Re}(H(\omega)\hat{P}_0(\omega)) \le -c_1 H(\omega)$;
3. $|H(\omega) - I| \le C_1/|\omega|$.

Under Assumptions 4a and 4b, a symmetrizer $H(\omega)$ satisfying the above Assumptions 4c1 and 4c2 can always be constructed by the Kreiss' matrix theorem [3]. Assumption 4c3 adds an additional restriction, which is fulfilled, however, for a large class of operators P_0, for which $u_t = P_0 u$ is hyperbolic with suitable zero-order term or parabolic or coupled parabolic-hyperbolic. We show this in the appendix. For generalizations, where Assumption 4c3 is dropped, see section 4.2.

For the coefficients $B_j(x,t),\ g_j(x,t,u),\ F_j(x,t)$ of (1.1) we make Assumption 3 and a symmetry assumption.

Assumption 5. The coefficients $B_j,\ g_j,\ F_j$ satisfy Assumption 3 and

$$B_j(x,t) = B_j^T(x,t), \quad g_{ju}(x,t,u) = g_{ju}^T(x,t,u)$$
$$\text{for all} \quad x \in \mathbb{R}^d,\ t \ge 0,\ u \in \mathbb{R}^n.$$

First consider (1.1) with $\varepsilon_2 = 0$; i.e., consider the linear equation

$$u_t = P_0 u + \varepsilon_1 \sum_j D_j(B_j u) + \sum_j D_j F_j. \tag{3.5}$$

If u is a solution, we set

$$G_j(x,t) = \varepsilon_1 B_j(x,t)u(x,t) + F_j(x,t)$$

and obtain

$$u_t = P_0 u + \sum_j D_j G_j. \tag{3.6}$$

Then Fourier–Laplace transformation yields

$$(sI - \hat{P}_0)\tilde{u} = i\sum_j \omega_j \tilde{G}_j, \quad s = \eta + i\xi,\ \eta \ge 0. \tag{3.7}$$

For $|\omega| \le 1$, our estimates are based on the following bound of $(sI - \hat{P}_0(\omega))^{-1}$.

LEMMA 3.1. *Consider a symbol $\hat{P}_0(\omega)$ satisfying Assumptions* 4a,b. *There is $C > 0$ so that*

$$|\omega|^2 \int_{-\infty}^{\infty} \left| \left((\eta + i\xi)I - \hat{P}_0 \right)^{-1} \right|^2 d\xi \le C \quad for\ all \quad |\omega| \le 1,\ \eta \ge 0.$$

Proof. By the Kreiss' matrix theorem (see [3]), there is a transformation $S = S(\omega)$ with

$$S^{-1}\hat{P}_0 S = \begin{bmatrix} \lambda_1 & r_{12} & \cdots & r_{1n} \\ & \ddots & \ddots & \vdots \\ & & \ddots & r_{n-1,n} \\ 0 & & & \lambda_n \end{bmatrix},$$

$$|r_{jl}| \le C_1 |\mathrm{Re}\ \lambda_j|, \quad 1 \le j < l \le n, \quad |S| + |S^{-1}| \le C_1.$$

If $\lambda_j = \alpha_j + i\beta_j,\ \alpha_j, \beta_j \in \mathbb{R}$, then we have $\alpha_j \le -c_0|\omega|^2$ by Assumption 4b, and one obtains

$$\left| \left((\eta + i\xi)I - \hat{P}_0 \right)^{-1} \right|^2 \le C_2 \sum_j \frac{1}{(\eta - \alpha_j)^2 + (\xi - \beta_j)^2}\ .$$

Using the bound

$$\eta - \alpha_j \ge c_0 |\omega|^2,$$

one finds that

$$\int_{-\infty}^{\infty} \frac{d\xi}{(\eta - \alpha_j)^2 + (\xi - \beta_j)^2} \le C_3 |\omega|^{-2}.$$

This proves the lemma.

Recall (3.6) and recall the decomposition $u = u^I + u^{II}$ determined by (1.5). We obtain the following estimate of u^I in terms of G.

LEMMA 3.2. *Let u solve* (3.6) *and let $u = 0$ at $t = 0$; let $\hat{P}_0$ satisfy Assumptions* 4a,b. *Then, for all $p = 0, 1, \ldots$ there is R_p, independent of T and G, with*

$$\int_0^T \|u^I\|_{H^p}^2 dt \le R_p \left\{ M(G, T) + \int_0^T \|G\|^2 dt \right\}. \tag{3.8}$$

Proof. From (3.7) we obtain

$$|\tilde{u}|^2 \le |\omega|^2 |(sI - \hat{P}_0)^{-1}|^2 |\tilde{G}|^2 \tag{3.9}$$

with

$$|\tilde{G}|^2 = \sum_j |\tilde{G}_j|^2 \le M(G, \infty).$$

By Parseval's relation,

$$\int_0^\infty e^{-2\eta t} \|u^I\|^2 dt = \frac{1}{2\pi} \int_{|\omega| \le 1} \int_{-\infty}^{\infty} |\tilde{u}(\omega, \eta + i\xi)|^2 d\xi\, d\omega.$$

Using estimate (3.9) and the previous lemma, we obtain

$$\int_0^\infty \|u^I\|^2 dt \le C\ M(G,\infty).$$

Estimates for $D^\alpha u^I$ are then obtained in the same way as in the proof of Theorem 2.1. Since values $G(x,t)$ for $t > T$ do not change the solution $u(x,t)$ for $t \le T$, we may replace $T = \infty$ by any finite T. This proves the lemma.

Now recall our assumption that u solves the linear equation (3.5), and we have defined

$$G_j = \varepsilon_1 B_j u + F_j, \quad j = 1, \dots, d$$

to derive (3.6). Let us denote

$$|||\ B|||^{\ 2} = \int_0^\infty \|B(\cdot,t)\|^2 dt$$

and

$$|B|_\infty = \sup\big\{|B(x,t)|\ :\ x \in \mathbb{R}^d,\ t \ge 0\big\}.$$

We want to replace G on the right side of (3.8) by $G = \varepsilon_1 Bu + F$. First note

$$\Big(\int_0^\infty \int_{\mathbb{R}^d} |G_j| dx\, dt\Big)^2 \le 2\varepsilon_1^2 \Big(\int_0^\infty \int_{\mathbb{R}^d} |B_j|\, |u| dx\, dt\Big)^2 + 2\Big(\int_0^\infty \int_{\mathbb{R}^d} |F_j| dx\, dt\Big)^2$$

$$\le 2\varepsilon_1^2 |||\ B|||^{\ 2} \int_0^\infty \|u\|^2 dt + 2M(F,\infty).$$

Furthermore,

$$\int_0^\infty \|G\|^2 dt \le 2\varepsilon_1^2 |B|_\infty^2 \int_0^\infty \|u\|^2 dt + 2\int_0^\infty \|F\|^2 dt.$$

Substituting these estimates into (3.8), one obtains the following result.

LEMMA 3.3. *Let u solve (3.5), $u = 0$ at $t = 0$. For all $p = 0, 1, \dots$ there is R_p, depending only on p and P_0, such that*

$$\int_0^T \|u^I\|_{H^p}^2 dt \le R_p \left\{ M(F,T) + \int_0^T \|F\|^2 dt \right\} + \varepsilon_1^2 R_p C_B \int_0^T \|u^{II}\|^2 dt.$$

Here $C_B = |||\ B|||^2 + |B|_\infty^2$, and it is assumed that

$$\varepsilon_1^2 R_p C_B \le \frac{1}{4}.$$

Now we estimate the u^{II}-part of the solution u of (3.5). Recall the properties of the symmetrizer $H(\omega)$ for $|\omega| \ge 1$, formulated in Assumption 4c. We set

$$H(\omega) = I \quad \text{for} \quad |\omega| < 1$$

and introduce a new inner product on $L_2 = L_2(\mathbb{R}^d; \mathbb{R}^n)$ by

$$(u, v)_{\mathcal{H}} = \int_{\mathbb{R}^d} \hat{u}^*(\omega) H(\omega) \hat{v}(\omega) d\omega.$$

The corresponding norm is

$$\|u\|_{\mathcal{H}} = (u, u)_{\mathcal{H}}^{1/2}.$$

The following result is an immediate consequence of the properties of $H(\omega)$ and Parseval's relation.

LEMMA 3.4.

a) $\frac{1}{C_1}\|u\|^2 \le \|u\|_{\mathcal{H}}^2 \le C_1\|u\|^2 \quad for\ all\ u \in L_2$;

b) $Re\,(u^{II}, P_0 u^{II})_{\mathcal{H}} \le -c_1 \|u^{II}\|_{\mathcal{H}}^2 \quad for\ all\ u \in H^m$;

c) $|(u, D_j v) - (u, D_j v)_{\mathcal{H}}| \le \text{const.}\|u\|\ \|v\| \quad for\ all\ u \in L_2,\ v \in H^1,\ j = 1, \dots, d.$

Our basic energy estimate for u^{II} is the following.

LEMMA 3.5. *Let u solve* (3.5) *and recall Assumptions* 4c, 5. *Then we have*

$$\frac{d}{dt}\|u^{II}\|_{\mathcal{H}}^2 \le -\frac{3}{2} c_1 \|u^{II}\|_{\mathcal{H}}^2 + C_0 \|F\|_{H^1}^2 + \varepsilon_1^2 C_B \|u^I\|^2$$

for $|\varepsilon_1| \le \varepsilon_0(B)$. The constant $c_1 > 0$ is the same as the constant c_1 in Lemma 4b*; the constant C_0 depends only on P_0; the constants $\varepsilon_0(B) > 0$ and C_B depend on $|B|_\infty + |DB|_\infty$, where*

$$|B|_\infty = \sup\big\{|B(x,t)| \ :\ x \in \mathbb{R}^d,\ t \ge 0\big\},$$
$$|DB|_\infty = \max_j \ \sup\Big\{|D_j B_j(x,t)| \ :\ x \in \mathbb{R}^d,\ t \ge 0\Big\}.$$

Proof. We have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|u^{II}\|_{\mathcal{H}}^2 &= \operatorname{Re}(u^{II}, u_t^{II})_{\mathcal{H}} \\ &= \operatorname{Re}(u^{II}, u_t)_{\mathcal{H}} \\ &= \operatorname{Re}(u^{II}, P_0 u^{II})_{\mathcal{H}} \\ &\quad + \sum_j \operatorname{Re}(u^{II}, D_j F_j)_{\mathcal{H}}.\end{aligned}$$

Here

$$\operatorname{Re}(u^{II}, P_0 u^{II})_{\mathcal{H}} \le -c_1 \|u^{II}\|_{\mathcal{H}}^2,$$

by Lemma 3.4b. The remaining terms on the right side of the above equality are estimated as follows.

First note

$$\begin{aligned}|(u^{II}, D_j F_j)_{\mathcal{H}}| &\le \text{ const. } \|u^{II}\|\ \|F\|_{H^1} \\ &\le \mu \|u^{II}\|^2 + \frac{C}{\mu}\|F\|_{H^1}^2\end{aligned}$$

for any $\mu > 0$. The term

$$\mathrm{Re}\,(u^{II}, D_j(B_j u))_{\mathcal{H}}, \quad u = u^I + u^{II},$$

is estimated in two parts. Since $\|D_j u^I\| \le \|u^I\|$, we have that

$$\begin{aligned}\left|\left(u^{II}, D_j(B_j u^I)\right)_{\mathcal{H}}\right| &\le C\|u^{II}\|\Big\{|B|_\infty + |DB|_\infty\Big\}\|u^I\| \\ &\le \mu\|u^{II}\|^2 + \frac{C_B}{\mu}\|u^I\|^2 \quad \text{for any } \mu > 0.\end{aligned}$$

Furthermore, to estimate

$$T_j := \mathrm{Re}\,(u^{II}, D_j(B_j u^{II}))_{\mathcal{H}}$$

we apply Lemma 3.4c to obtain

$$\begin{aligned}T_j &= \mathrm{Re}\,(u^{II}, D_j(B_j u^{II})) + T_j', \\ |T_j'| &\le C|B|_\infty\|u^{II}\|^2.\end{aligned}$$

Using integration by parts and the symmetry of B_j we also have

$$\left|\mathrm{Re}\left(u^{II}, D_j(B_j u^{II})\right)\right| \le \frac{1}{2}|DB|_\infty\|u^{II}\|^2.$$

To summarize, we have shown that

$$\begin{aligned}\frac{d}{dt}\|u^{II}\|_{\mathcal{H}}^2 \le -2c_1\|u^{II}\|_{\mathcal{H}}^2 + +\mu\|u^{II}\|^2 + \frac{1}{\mu}C\|F\|_{H^1}^2 \\ + \mu\|u^{II}\|^2 + \frac{\varepsilon_1^2}{\mu}C_B\,\|u^I\|^2 + |\varepsilon_1|C_B\,\|u^{II}\|^2.\end{aligned}$$

Choosing $\mu > 0$ sufficiently small, the result follows.

The basic energy estimate formulated in the previous lemma will now be used to derive similar estimates for space derivatives $D^\alpha u^{II}$. Applying D^α to (3.5), we obtain

$$D^\alpha u_t = P_0 D^\alpha u + \varepsilon_1 \sum_j D_j(B_j D^\alpha u) + \varepsilon_1 \sum_j D_j\Phi_j + \sum_j D_j D^\alpha F_j$$

with[1]

$$\Phi_j = D^\alpha(B_j u) - B_j(D^\alpha u) = \sum_{\beta<\alpha} c_{\alpha\beta}(D^{\alpha-\beta}B_j)D^\beta u. \tag{3.10}$$

Let us define

$$\|u\|_{p,\mathcal{H}}^2 = \sum_{|\alpha|\le p}\|D^\alpha u\|_{\mathcal{H}}^2.$$

[1]We use the notation $\beta < \alpha$ to mean $\beta_j \le \alpha_j$ for $j = 1, \dots, d$ with at least one strict inequality.

We will show the following generalization of the previous lemma.

LEMMA 3.6. *Let u solve* (3.5) *and recall Assumptions* 4c, 5. *For any $p = 0, 1, \ldots$ we have*

$$\frac{d}{dt}\|u^{II}\|^2_{p,\mathcal{H}} \le -c_1\|u^{II}\|^2_{p,\mathcal{H}} + C_p\|F\|^2_{H^{p+1}} + \varepsilon_1^2 C(B,p)\|u^I\|^2$$

for $|\varepsilon_1| \le \varepsilon_0(B,p)$. The constant $c_1 > 0$ is the same as the constant c_1 in Lemma 4b*; the constant C_p depends only on p and P_0; the constants $\varepsilon_0(B,p) > 0$ and $C(B,p)$ are independent of F.*

Proof. Let $|\alpha| \le p$. We apply Lemma 3.5 (with F replaced by $D^\alpha F + \varepsilon_1\Phi$) to obtain

$$\frac{d}{dt}\|D^\alpha u^{II}\|^2_{\mathcal{H}} \le -\frac{3c_1}{2}\|D^\alpha u^{II}\|^2_{\mathcal{H}} + C_1\Big\{\|D^\alpha F\|^2_{H^1} + \varepsilon_1^2\|\Phi\|^2_{H^1}\Big\} + \varepsilon_1^2 C_B\,\|u^I\|^2.$$

(Note that $\|D^\alpha u^I\| \le \|u^I\|$.) The main point is that Φ_j only contains derivatives of u of order $\le p-1$, and the coefficients $D^{\alpha-\beta}B_j$ in (3.10) are uniformly bounded. Therefore,

$$\|\Phi\|_{H^1} \le C(B,p)\|u\|_{H^p}.$$

Adding the resulting estimates of $(d/dt)\|D^\alpha u^{II}\|^2_{\mathcal{H}}$ for $|\alpha| \le p$, the lemma follows.

We now combine the estimates derived in Lemma 3.3 for u^I and in Lemma 3.6 for u^{II} and prove the following result.

THEOREM 3.7. *Consider the linear problem* (3.5) *with $u = 0$ at $t = 0$ and recall Assumptions* 4 *and* 5. *For any $p = 0, 1, \ldots$ there are positive constants $\varepsilon_0 = \varepsilon_0(B,p) > 0$ and R_p so that*

$$\int_0^\infty \Big\{\|u\|^2_{H^{p+m}} + \|u_t\|^2_{H^p}\Big\}dt \le R_p\Big\{M(F,\infty)dt\Big\}$$

if $|\varepsilon_1| \le \varepsilon_0(B,p)$. The constant R_p depends only on p and P_0. (In the above estimate, m is the order of P_0, or $m = 1$ if P_0 has order zero.) Consequently,

$$\lim_{t\to\infty} |u(\cdot,t)| = 0.$$

Proof. Let us abbreviate

$$\begin{aligned} y_1(t) &= \|u^I(\cdot,t)\|^2_{H^p},\\ y_2(t) &= \|u^{II}(\cdot,t)\|^2_{p,\mathcal{H}},\\ C_F &= M(F,\infty) + \int_0^\infty \|F(\cdot,t)\|^2 dt,\\ f(t) &= \|F(\cdot,t)\|^2_{H^{p+1}}. \end{aligned}$$

Then we have shown that there are positive constants $C = C(B,p), R_p$, and $\varepsilon_0 = \varepsilon_0(B,p)$ so that, for $|\varepsilon_1| \le \varepsilon_0$,

$$\begin{aligned} &\int_0^\infty y_1 dt \le R_p C_F + \varepsilon_1^2 C\int_0^\infty y_2 dt,\\ &\frac{d}{dt}y_2 \le -c_1 y_2 + R_p f + \varepsilon_1^2 C y_1. \end{aligned} \tag{3.11}$$

Since $y_2(0) = 0$ we obtain that

$$y_2(t) \le R_p \int_0^t e^{-c_1(t-\tau)} f(\tau) d\tau + \varepsilon_1^2 C \int_0^t e^{-c_1(t-\tau)} y_1(\tau) d\tau,$$

and, therefore,

$$\int_0^\infty y_2(t) dt \le R_p \int_0^\infty \int_0^t e^{-c_1(t-\tau)} f(\tau) d\tau dt + \varepsilon_1^2 C \int_0^\infty \int_0^t e^{-c_1(t-\tau)} y_1(\tau) d\tau dt.$$

Clearly,

$$\begin{aligned} \int_0^\infty \int_0^t e^{-c_1(t-\tau)} f(\tau) d\tau dt &= \int_0^\infty \int_\tau^\infty e^{-c_1(t-\tau)} f(\tau) dt d\tau \\ &\le \frac{1}{c_1} \int_0^\infty f(\tau) d\tau, \end{aligned}$$

and in the same way

$$\int_0^\infty \int_0^t e^{-c_1(t-\tau)} y_1(\tau) d\tau dt \le \frac{1}{c_1} \int_0^\infty y_1(\tau) d\tau.$$

This proves the estimate

$$\int_0^\infty y_2 dt \le \frac{R_p}{c_1} \int_0^\infty f dt + \varepsilon_1^2 \frac{C}{c_1} \int_0^\infty y_1 dt. \tag{3.12}$$

By adding the inequalities (3.11) and (3.12) and using the equivalence of the norms $\|\cdot\|_{H^p}$ and $\|\cdot\|_{p,\mathcal{H}}$, we have shown

$$\int_0^\infty \|u\|_{H^p}^2 dt \le R_p' \left\{ M(F, \infty) + \int_0^\infty \|F\|_{H^{p+1}}^2 dt \right\} \tag{3.13}$$

for $|\varepsilon_1| \le \varepsilon_0(B, p)$. To estimate $D^\alpha u_t$, we use the differential equation (3.5) to replace u_t. This gives us

$$\|u_t\|_{H^p}^2 \le C \Big\{ \|u\|_{H^{p+m}}^2 + \|F\|_{H^{p+1}}^2 \Big\}.$$

Integrating in time and applying (3.13) with p replaced by $p + m$, we obtain the desired estimate. Convergence of u to zero as $t \to \infty$ follows as before when $p > d/2$. This finishes the proof of the theorem.

Now consider the nonlinear problem (1.1), (1.2). By combining the techniques to prove Theorems 2 and 3, we show nonlinear stability.

THEOREM 3.8. *Consider the nonlinear problem* (1.1) *with* $u = 0$ *at* $t = 0$ *and recall Assumptions* 4 *and* 5. *There exists* $\varepsilon_0 > 0$ *so that the solution is* C^∞ *if* $\varepsilon_1^2 + \varepsilon_2^2 \le \varepsilon_0$. *Furthermore,*

$$\lim_{t \to \infty} |u(\cdot, t)| = 0.$$

Proof. Let κ denote a large constant to be determined below. Fix $\varepsilon_1, \varepsilon_2$ and let $u = u(x,t,\varepsilon_1,\varepsilon_2)$ denote the solution of (1.1), (1.2). Suppose there exists $T = T(\kappa,\varepsilon_1,\varepsilon_2)$ with

$$\int_0^T \Big\{ \|u\|^2_{H^{p+m}} + \|u_t\|^2_{H^p} \Big\} dt = \kappa. \tag{3.14}$$

Here $p = m + d + 2$; the number m is the order of P_0, or $m = 1$ if P_0 has order zero. In the following, C denotes a constant independent of $\varepsilon_1, \varepsilon_2, T, \kappa$; also, C_κ denotes a constant independent of $\varepsilon_1, \varepsilon_2$, and T, which may, however, depend on κ. The values of C and C_κ may change at different occurrences.

By Sobolev's inequality, there is C such that

$$\|D^\alpha u\|^2_{\infty,T} \le C\kappa \quad \text{if} \quad |\alpha| + \frac{d}{2} < p.$$

We abbreviate

$$\begin{aligned} y_1(t) &= \|u^I(\cdot,t)\|^2_{H^{p+m}}, \\ y_2(t) &= \|u^{II}(\cdot,t)\|^2_{p+m,\mathcal{H}}, \\ C_F &= M(F,\infty) + \int_0^\infty \|F(\cdot,t)\|^2 dt, \\ f(t) &= \|F(\cdot,t)\|^2_{H^{p+m+1}}. \end{aligned}$$

To estimate u^I, note that u satisfies

$$u_t = P_0 u + \sum_j D_j Y_j$$

with

$$Y_j = F_j + \varepsilon_1 B_j u + \varepsilon_2 G_j, \quad G_j(x,t) = g_j(x,t,u(x,t)).$$

We obtain by Lemma 2,

$$\begin{aligned} \int_0^T y_1 dt &\le R_{p+m} \Big\{ M(F + \varepsilon_1 Bu + \varepsilon_2 G, T) + \int_0^T \|F + \varepsilon_1 Bu + \varepsilon_2 G\|^2 dt \Big\} \\ &\le C C_F + (\varepsilon_1^2 + \varepsilon_2^2) C_\kappa. \end{aligned} \tag{3.15}$$

The last estimate follows as in the proof of Theorem 2.

To estimate u^{II}, we use $\|\cdot\|_{\mathcal{H}}$. For $1 \le |\alpha| \le p+m$ the function $D^\alpha u$ satisfies

$$\begin{aligned} D^\alpha u_t = P_0 D^\alpha u &+ \varepsilon_1 \sum_j D_j (B_j D^\alpha u) \\ &+ \varepsilon_2 \sum_j D_j (g_{ju} D^\alpha u) + \varepsilon_1 \sum_j D_j \Phi_j \\ &+ \varepsilon_2 \sum_j D_j \Psi_j + \sum D_j D^\alpha F_j. \end{aligned}$$

Here Φ_j is given in (3.10) and

$$\Psi_j = D^\alpha G_j - g_{ju} D^\alpha u$$

is a sum of terms

$$\psi_j(x,t,\alpha,\sigma) D^{\sigma_1} u \cdots D^{\sigma_r} u$$

where

$$|\sigma_1| + \cdots + |\sigma_r| \le p+m, \quad |\sigma_j| < p+m,\ j = 1, \ldots, r.$$

The function $\psi_j(x,t,\alpha,\sigma)$ is a derivative $D_x^\beta D_u^\gamma g_j$ evaluated at $(x,t,u(x,t))$. To estimate $D^\alpha u^{II}$ we apply Lemma 5 with $\varepsilon_1 B_j$ replaced by $\varepsilon_1 B_j + \varepsilon_2 g_{ju}$, and F replaced by $D^\alpha F + \varepsilon_1 \Phi + \varepsilon_2 \Psi$. This yields

$$\frac{d}{dt}\|D^\alpha u^{II}\|_{\mathcal{H}}^2 \le -\frac{3}{2} c_1 \|D^\alpha u^{II}\|_{\mathcal{H}}^2 + C\|D^\alpha F + \varepsilon_1 \Phi + \varepsilon_2 \Psi\|_{H^1}^2 + (\varepsilon_1^2 + \varepsilon_2^2) C_\kappa \|u^I\|^2.$$

Here $\|\Phi\|_{H^1} \le C\|u\|_{H^{p+m}}$ and $\|\Psi\|_{H^1} \le C_\kappa \|u\|_{H^{p+m}}$. The estimate of Ψ follows as in the proof of Theorem 2. (The main point is that all derivatives of u occurring in Ψ_j are of order strictly less than $p+m$. Also, $D_j \Psi_j$ is a sum of terms

$$\rho_j(x,t,\alpha,\sigma) D^{\sigma_1} u \cdots D^{\sigma_r} u$$

with

$$|\sigma_1| + \cdots + |\sigma_r| \le p+m+1.$$

If there were two factors, $D^{\sigma_1} u$ and $D^{\sigma_2} u$, say, which cannot be estimated in sup-norm in terms of κ, then

$$|\sigma_1| + \frac{d}{2} \ge p, \quad |\sigma_2| + \frac{d}{2} \ge p;$$

thus

$$p + m + 1 + d \ge |\sigma_1| + |\sigma_2| + d \ge 2p.$$

This contradicts our choice of $p = m + d + 2$.)

An estimate for

$$\frac{d}{dt}\|u^{II}\|_{\mathcal{H}}^2$$

(where u^{II} is undifferentiated), can also be obtained from Lemma 5 with $\varepsilon_2 g_j$ included as forcing term. Adding the resulting estimates for $|\alpha| \le p+m$, one obtains

$$\frac{d}{dt} y_2 \le -c_1 y_2 + Cf + (\varepsilon_1^2 + \varepsilon_2^2) C_\kappa \|u^I\|^2$$

for $\varepsilon_1^2 + \varepsilon_2^2 \le \varepsilon_0(\kappa)$. As in the proof of Theorem 3 we find

$$\int_0^T y_2 dt \le C \int_0^T f dt + (\varepsilon_1^2 + \varepsilon_2^2) C_\kappa. \tag{3.16}$$

Adding the estimates (3.15) and (3.16), we have shown that

$$\int_0^T \|u\|^2_{H^{p+m}} dt \le C\left(C_F + \int_0^T f dt\right) + (\varepsilon_1^2+\varepsilon_2^2)C_\kappa \tag{3.17}$$

for $\varepsilon_1^2+\varepsilon_2^2 \le \varepsilon_0(\kappa)$. We now use the differential equation (1.1) to estimate the H^p-norm of u_t. This yields

$$\|u_t\|^2_{H^p} \le C\|u\|^2_{H^{p+m}} + C\|F\|^2_{H^{p+1}} + (\varepsilon_1^2+\varepsilon_2^2)C_\kappa \|u\|^2_{H^{p+1}}.$$

Integrating in time and using (3.17) one finds

$$\int_0^T \left\{\|u\|^2_{H^{p+m}} + \|u_t\|^2_{H^p}\right\} dt \le C\left\{M(F,\infty) + \int_0^\infty \|F\|^2_{H^{p+m+1}} dt\right\} + (\varepsilon_1^2+\varepsilon_2^2)C_\kappa. \tag{3.18}$$

Therefore, we choose

$$\kappa = 1 + C\left\{M(F,\infty) + \int_0^\infty \|F\|^2_{H^{p+m+1}} dt\right\}$$

and let

$$(\varepsilon_1^2+\varepsilon_2^2)C_\kappa \le \frac{1}{2}.$$

Then the estimate (3.18) shows that a finite T with (3.14) does not exist. The remaining arguments are the same as in the proof of Theorem 2.

Remark. The more direct approach to estimate the time derivative u_t, which we have used in section 2, does not seem possible under the assumptions of section 3.

4. Discussion, conjecture.

4.1. Discussion of Assumption 3. To illustrate our assumption of the specific form of the Cauchy problem (1.1), (1.2) and our general Assumption 3 for $B_j(x,t), g_j(x,t,u), F_j(x,t)$, we consider a system

$$u_t = P_0 u + \sum_j D_j f_j(u), \tag{4.1}$$

where P_0 has constant coefficients and the $f_j : \mathbb{R}^n \to \mathbb{R}^n$ are smooth functions vanishing quadratically at $u = 0$. Thus, $u \equiv 0$ is a stationary solution. We consider (4.1) with small initial data,

$$u(x,0) = \varepsilon U_0(x), \quad U_0 \in C^\infty.$$

Requirements for U_0 will be derived below. For simplicity of presentation, let us assume that $f_j(u)$ is quadratic in u, i.e.,

$$f_j(u) = Q_j(u,u),$$

where $Q_j : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is bilinear. If we write $u(x,t) = \varepsilon v(x,t)$, then (4.1) becomes

$$v_t = P_0 v + \varepsilon \sum_j D_j f_j(v), \quad v(x,0) = U_0(x). \tag{4.2}$$

To enforce homogeneous initial data and Assumption 3b, we set

$$\bar{v}(x,t) = e^{-t}(U_0(x) + tV_0(x)),$$

where V_0 will be determined below. Then we define a new variable $w(x,t)$ by $v = \bar{v} + w$ and obtain from (4.2),

$$w_t = P_0 w + \varepsilon \sum_j D_j\big(Q_j(\bar{v}, w) + Q_j(w, \bar{v})\big) + \varepsilon \sum_j D_j f_j(w) + E(x,t)$$

with the forcing function

$$E(x,t) = P_0\bar{v} - \bar{v}_t + \varepsilon \sum_j D_j f_j(\bar{v}).$$

At $t = 0$ we have

$$E(x,0) = P_0 U_0 + U_0 - V_0 + \varepsilon \sum_j D_j f_j(U_0).$$

Thus, we can enforce $E(x,0) \equiv 0$ by a proper choice of V_0. Also, if the initial function $U_0(x)$ has the form

$$U_0(x) = \sum_j D_j U_{0j}(x), \quad U_{0j} \in C^\infty,$$

and we construct $V_0 = \sum_j D_j V_{0j}$ correspondingly, then

$$E = \sum_j D_j F_j(x,t) \quad \text{with} \quad F_j(x,0) \equiv 0.$$

It is not difficult to verify Assumption 3 for the resulting system for w if, for example,

$$D^\alpha U_{0j} \in L_1 \cap L_\infty, \quad j = 1, \dots, d,$$

for all α.

4.2. Conjecture. We assume here that $u_t = P_0 u$ is a coupled parabolic-hyperbolic system and write P_0 in the form

$$P_0 = \sum_{j,l=1}^{d} B_{jl} D_j D_l + \sum_{j=1}^{d} A_j D_j + L$$

with $B_{j,l}, A_j, L \in \mathbb{R}^{n\times n}$. The symbol of P_0 is

$$\hat{P}_0(\omega) = -B(\omega) + iA(\omega) + L$$

with

$$B(\omega) = \sum_{j,l=1}^{d} B_{jl}\omega_j\omega_l, \quad A(\omega) = \sum_{j=1}^{d} A_j\omega_j.$$

If $u_t = P_0 u$ is not parabolic, then the results of section 2 do not apply, and the results of section 3 require symmetry of $B_j(x,t)$ and $g_{ju}(x,t,u)$. We want to relax the symmetry requirement and formulate the following conditions on $\hat{P}_0(\omega)$.

Assumption 6.
a) There are positive constants C_0, c_0 such that

$$|e^{\hat{P}_0(\omega)t}| \le C_0 \quad \operatorname{Re}\lambda \le -c_0 \frac{|\omega|^2}{1+|\omega|^2} \text{ for all } \lambda \in \sigma(\hat{P}_0(\omega)).$$

b) There is a constant $c_1 > 0$ and, for every ω with $|\omega| = 1$, there is a nonsingular transformation $T(\omega)$ such that

$$T^{-1}(\omega)B(\omega)T(\omega) = \begin{pmatrix} B_1(\omega) & 0 \\ 0 & 0 \end{pmatrix}, \quad B_1 + B_1^* \ge 2c_1 I.$$

c) For $|\omega| = 1$, write

$$T^{-1}(\omega)A(\omega)T(\omega) = \begin{pmatrix} A_{11}(\omega) & A_{12}(\omega) \\ A_{21}(\omega) & A_{22}(\omega) \end{pmatrix},$$

where $T^{-1}AT$ has the same block structure as $T^{-1}BT$. Then the eigenvalues of $A_{22}(\omega)$ are real and distinct. *(*$A_{22}(\omega)$ defines a strictly hyperbolic pseudodifferential operator.)

We conjecture that the same conclusion as formulated in Theorem 2 is valid for the nonlinear problem (1.1), (1.2) if we make our general Assumption 3 for B_j, g_j, F_j. Thus, in contrast to our assumptions in section 3, we do *not* require *symmetry* of B_j, g_{ju} to obtain nonlinear stability.

We also conjecture that the assumption of strict hyperbolicity can be weakened as follows. If we set

$$\Psi(x,t,u,\omega,\varepsilon_1,\varepsilon_2) = \varepsilon_1 \sum_j \omega_j B_j(x,t) + \varepsilon_2 \sum_j \omega_j g_{ju}(x,t,u)$$

and, similarly as above, define the block Ψ_{22} by

$$T^{-1}(\omega)\Psi T(\omega) = \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix},$$

then we require that all eigenvalues of

$$A_{22}(\omega) + \Psi_{22}(x,t,u,\omega,\varepsilon_1,\varepsilon_2)$$

are real and have constant multiplicity for $\varepsilon_1^2 + \varepsilon_2^2$ small.

5. Appendix. Consider a constant coefficient operator

$$P = \sum_{j,l=1}^{d} B_{jl} D_j D_l + \sum_{j=1}^{d} A_j D_j + L \tag{5.1}$$

with $B_{jl}, A_j, L \in \mathbb{C}^{n\times n}$. The symbol of P is

$$\hat{P} = \hat{P}(\omega) = -B(\omega) + iA(\omega) + L, \quad \omega \in \mathbb{R}^d,$$

with

$$B(\omega) = \sum_{j,l=1}^{d} B_{jl}\omega_j\omega_l, \quad A(\omega) = \sum_{j=1}^{d} A_j\omega_j.$$

We make the following assumptions.

Assumption 7.

a) $A_j = A_j^*, \quad j = 1, \dots, d;$

b) for all $\omega^0 \in \mathbb{R}^d$ with $|\omega^0| = 1$ there is a unitary transformation $U = U(\omega^0)$ with

$$U^* B(\omega^0) U = \begin{bmatrix} B_1(\omega^0) & 0 \\ 0 & 0 \end{bmatrix}, \quad B_1 + B_1^* > 0;$$

c) there is a unitary transformation V with

$$V^* L V = \begin{bmatrix} L_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad L_1 + L_1^* < 0;$$

d) for all $\omega^0 \in \mathbb{R}^d$ with $|\omega^0| = 1$ and all $\varphi \in \mathbb{C}^n$, the following holds. If

$$B(\omega^0)\varphi = L\varphi = 0,$$

then φ is not an eigenvector of $A(\omega_0)$.

Remarks. 1) Assumption 7b allows for $B(\omega^0) = 0$; the condition $B_1 + B_1^* > 0$ is only required if the block B_1 is not empty. A similar remark applies to L. In particular, the conditions allow for $B(\omega) \equiv 0$ (the hyperbolic case) or $L = 0$ (the case without zero order term).

2) For $L = 0$, condition 7d is called *interaction condition* in [1] since it requires interaction of the characteristic variables of the symmetric hyperbolic system $u_t = \sum_j A_j D_j u$ with the parabolic part of the whole system.

3) For real symmetric B_{jl} and L, the conditions are also used in [6].

THEOREM 5.1. *If P has the form* (5.1) *and satisfies Assumption* 7*, then P satisfies Assumption* 4.

Proof. 1) First fix $\omega^0 \in \mathbb{R}^d$ with $|\omega^0| = 1$ and abbreviate

$$A = \sum_{j=1}^{d} A_j \omega_j^0, \quad B = \sum_{j,l=1}^{d} B_{jl} \omega_j^0 \omega_l^0,$$
$$\hat{P} = \hat{P}(\omega^0) = -B + iA + L.$$

Let

$$\hat{P}\varphi = \lambda\varphi, \quad |\varphi| = 1.$$

Then we have

$$\varphi^*(\hat{P} + \hat{P}^*)\varphi = 2 \operatorname{Re} \lambda,$$

and, since

$$\hat{P} + \hat{P}^* = -(B + B^*) + L + L^* \le 0,$$

it follows that $\operatorname{Re} \lambda \le 0$.

Suppose that $\operatorname{Re} \lambda = 0$; then

$$\varphi^*(B + B^*)\varphi = \varphi^*(L + L^*)\varphi = 0.$$

This would imply $B\varphi = L\varphi = 0$. Since $\hat{P}\varphi = \lambda\varphi$, it would follow that φ is an eigenvector of A, in contradiction to Assumption 7d. Therefore, $\operatorname{Re}\lambda < 0$ for all eigenvalues of $\hat{P}(\omega^0),\ |\omega^0| = 1$.

2) Using a well-known construction[2] (see, e.g., [3]), one obtains that there are constants $c_0 > 0,\ C_0 > 0$ and, for every ω^0 with $|\omega^0| = 1$, there is a Hermitian matrix $H_0 = H_0(\omega^0)$ such that

$$H_0\hat{P}(i\omega^0) + \hat{P}^*(i\omega^0)H_0 \le -c_0 H_0, \tag{5.2}$$
$$0 < \frac{1}{C_0} I \le H_0 \le C_0 I.$$

The function $H_0(\omega^0)$ can be constructed as a C^∞-function on the unit sphere $|\omega^0| = 1$ in $\mathbb{R}^d$.

3) We now construct $H(\omega)$ for $|\omega| \ge 1$ as follows. Let

$$\omega = \varrho\omega^0, \quad |\omega^0| = 1, \quad \varrho = |\omega|,$$
$$\hat{P} = \hat{P}(\omega) = -\varrho^2 B + i\varrho A + L,$$
$$\hat{P}_0 = \hat{P}(\omega^0) = -B + iA + L.$$

We set

$$H(\omega) = I + \frac{\alpha}{\varrho} H_0(\omega^0),$$

where $\alpha > 0$ will be determined below as a sufficiently small constant. Then we have

$$\begin{aligned} H\hat{P} + \hat{P}^*H &= -\varrho^2(B + B^*) + L + L^* \\ &\quad -\alpha\varrho(H_0B + B^*H_0) + i\alpha(H_0A + A^*H_0) \\ &\quad +\frac{\alpha}{\varrho}(H_0L + L^*H_0) \\ &= \alpha\left\{H_0\hat{P}_0 + \hat{P}_0^*H_0\right\} + T_1 + T_2 \qquad (5.3) \\ &\le -\alpha c_0 H_0 + T_1 + T_2, \end{aligned}$$

where

$$T_1 = -\varrho^2(B + B^*) + (\alpha - \alpha\varrho)(H_0B + B^*H_0),$$
$$T_2 = L + L^* + \left(\frac{\alpha}{\varrho} - \alpha\right)(H_0L + L^*H_0),$$

and where (5.2) is used in the final estimate.

Assuming first that

$$B = \begin{bmatrix} B_1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_1 + B_1^* > 0, \tag{5.4}$$

and partitioning

$$\varphi = \begin{bmatrix} \varphi^I \\ \varphi^{II} \end{bmatrix}, \quad \varphi \in \mathbb{C}^n,$$

[2]Using Schur's theorem, there is a transformation $T = UD, D = diag(1, \varepsilon, \dots, \varepsilon^{n-1})$, so that $\operatorname{Re}(T^{-1}\hat{P}(\omega^0)T)) \le -c_0 I < 0$. Then $H_0 = (T^{-1})^*T^{-1}$ satisfies $\operatorname{Re}(H_0\hat{P}(\omega^0)) \le -c_0 H_0$. A partition of unity argument shows that one can construct $H_0(\omega^0)$ as C^∞-function on the unit sphere $|\omega^0| = 1$.

accordingly, one obtains

$$\begin{aligned}\varphi^* T_1 \varphi &\le -c_1 \varrho^2 |\varphi^I|^2 + C_1 \alpha \varrho |\varphi|\, |\varphi^I| \\ &\le -c_1 \varrho^2 |\varphi^I|^2 + \alpha \left\{ \mu |\varphi|^2 + C(\mu) \varrho^2 |\varphi^I|^2 \right\},\end{aligned}$$

for any $\mu > 0$. By first choosing $\mu > 0$ so small that

$$\mu |\varphi|^2 \le \frac{c_0}{4} \varphi^* H_0 \varphi$$

and then choosing $\alpha > 0$ so small that

$$\alpha C(\mu) \le c_1,$$

one obtains

$$\varphi^* T_1 \varphi \le \alpha \frac{c_0}{4} \varphi^* H_0 \varphi.$$

It is easy to see that Assumption (5.4) can be replaced by Assumption 7b. Estimating the term T_2 similarly, one obtains

$$H\hat{P} + \hat{P}^* H \le -\alpha \frac{c_0}{2} H_0 \le -\tilde{c}_0 H, \quad \tilde{c}_0 > 0. \tag{5.5}$$

This finishes the construction of $H(\omega)$ for $|\omega| \ge 1$ in Assumption 4c.

4) Now consider

$$\hat{P} = -\varrho^2 B + i\varrho A + L$$

for $0 < \varrho \le 1$. We set

$$H(\omega) = H(\varrho \omega^0) = I + \alpha \varrho H_0(\omega^0)$$

and obtain

$$\begin{aligned}H\hat{P} + \hat{P}^* H = \varrho^2 \Big\{ &- (B + B^*) + \varrho^{-2}(L + L^*) - \alpha\varrho (H_0 B + B^* H_0) \\ &+ i\alpha (H_0 A + A^* H_0) + \frac{\alpha}{\varrho} (H_0 L + L^* H_0) \Big\}.\end{aligned}$$

The term in the brackets $\{\dots\}$ agrees exactly with the right side of (5.3) if L and $(-B)$ are exchanged and ϱ is replaced by $1/\varrho$. Therefore, the estimate (5.5) implies

$$H\hat{P} + \hat{P}^* H \le -\tilde{c}_0 \varrho^2 H.$$

Consequently, if λ is an eigenvalue of $\hat{P}$, it follows that

$$2 \operatorname{Re} \lambda \le -\tilde{c}_0 \varrho^2.$$

Furthermore, the proven estimates

$$H\hat{P} + \hat{P}^* H \le 0,$$

together with the uniform bounds

$$0 < \frac{1}{C} I \le H \le CI,$$

imply

$$\left| e^{\hat{P}t} \right| \le K, \quad t \ge 0.$$

This finishes the proof of the theorem.

REFERENCES

[1] T. Hagstrom and J. Lorenz, *All-time existence of smooth solutions to PDEs of mixed type and the invariant subspace of uniform states*, Adv. Appl. Math., 16 (1995), pp. 219–257.
[2] S. Kawashima, *Large-time behaviour of solutions to hyperbolic-parabolic systems of conservation laws and applications*, in Proc. Roy. Soc. Edinburgh Sect. A, 106 (1987), pp. 169–194.
[3] H.-O. Kreiss and J. Lorenz, *Initial-boundary value problems and the Navier–Stokes equations*, Academic Press, New York, 1989.
[4] H.-O. Kreiss, O. Ortiz, and O. Reula, *Stability of quasi-linear hyperbolic dissipative systems*, J. Differential Equations, 142 (1998), pp. 78–96.
[5] R. Racke, *Lectures on Nonlinear Evolution Equations: Initial Value Problems*, Vieweg-Verlag, Braunschweig, Wiesbaden, 1992.
[6] Y. Shizuta and S. Kawashima, *Systems of equations of hyperbolic-parabolic type with applications to the discrete Boltzmann equation*, Hokkaido Math. J., 14 (1985), pp. 249–275.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 431–453

MATHEMATICAL ANALYSIS FOR RESERVOIR MODELS*

ZHANGXIN CHEN† AND RICHARD EWING‡

Abstract. In the first part of this paper, the mathematical analysis is presented in detail for the single-phase, miscible displacement of one fluid by another in a porous medium. It is shown that initial boundary value problems with various boundary conditions for this miscible displacement possess a weak solution under physically reasonable hypotheses on the data. In the second part of this paper, it is proven how the analysis can be extended to two-phase fluid flow and transport equations in a porous medium. The flow equations are written in a fractional flow formulation so that a degenerate elliptic-parabolic partial differential system is produced for a global pressure and a saturation. This degenerate system is coupled to a parabolic transport equation which models the concentration of one of the fluids. The analysis here does not utilize any regularized problem; a weak solution is obtained as a limit of solutions to discrete time problems.

Key words. porous medium, flow and transport, elliptic-parabolic system, degenerate equations, existence

AMS subject classifications. 35K60, 35K65, 76S05, 76T05

PII. S0036141097319152

1. Introduction.

Multiphase flow and transport of fluids in porous media is of importance socially and economically in a number of applications. For example, petroleum engineers have been interested in efficient recovery of energy resources, and hydrologists have been concerned with improvement of groundwater resource utilization for many years. Unfortunately, despite the great progress made in development of physical models of multiphase flow and transport of fluids in porous media, there has been limited mathematical theory behind these models. The difficulty stems from the fact that the equations modeling complex physical phenomena involving both flow and transport of fluids are often coupled systems of nonlinear, time-dependent degenerate partial differential equations. Hence, simplified models have been analyzed in the last 20 years.

The simplest porous media problem corresponds to the flow of a fluid where a whole porous medium is filled with the single fluid (usually gas or oil in petroleum engineering or water in groundwater hydrology). The usual equations for the single-phase flow model are of parabolic type for the fluid density or pressure and are well understood (see, e.g., [5, 8]).

A more complex case involves the single-phase, miscible displacement of one fluid by another in a porous medium. Under the assumption that no volume change results from the mixing of the two fluids, a coupled, nonlinear differential system of two equations is often utilized for this miscible displacement problem. One of the equations is of elliptic (respectively, parabolic) type for the fluid pressure if the fluids are incompressible (respectively, compressible), and the other is of parabolic type for the concentration of one fluid. This system is complicated by the facts that the pressure

*Received by the editors October 17, 1997; accepted for publication May 22, 1998; published electronically February 2, 1999. This research was partially supported by National Science Foundation grant DMS-9626179.

http://www.siam.org/journals/sima/30-2/31915.html

†Department of Mathematics, Box 750156, Southern Methodist University, Dallas, TX 75275-0156 (zchen@dragon.math.smu.edu).

‡Institute for Scientific Computation, Texas A&M University, College Station, TX 77843–3404 (ewing@ewing.tamu.edu).

equation can be degenerate due to the form of the concentration-dependent viscosity and that the transport and diffusion-dispersion coefficients in the concentration equation can be unbounded due to the potentially unbounded fluid velocity.

The miscible displacement problem was first studied by Sammon [20], where a one-dimensional model was theoretically analyzed and the viscosity of the fluid was assumed to be independent of the concentration. The latter assumption decouples the pressure equation from the concentration equation. Mikelić [18] later analyzed a three-dimensional stationary displacement problem. While the viscosity was allowed to be concentration-dependent, it was in fact assumed to be sufficiently close to a constant. Then the pressure and concentration equations were essentially separated so that well-posedness for the stationary problem can be established. Recently, Feng [14] considered a two-dimensional model for the displacement problem. The viscosity can be concentration-dependent, but the analysis was valid only for a two-dimensional problem. Furthermore, the analysis, following Kružkov and Sukorjanskiĭ [16] for two-dimensional two-phase immiscible flow, made use of the corresponding regularized system and required the coefficients of the regularized problem to be bounded uniformly with respect to the regularization parameter. The uniform boundedness is hardly satisfied due to the above mentioned feature for the transport and diffusion-dispersion coefficients. Finally, all the theoretical results in [20, 18, 14] were obtained solely for homogeneous Neumann boundary conditions and without gravity effects. The study of the single-phase, miscible displacement of one incompressible fluid by another from the numerical point of view using finite element methods has been extensively carried out (see, e.g., [12]); the case where the components are assumed to be slightly compressible also has been numerically studied (see, e.g., [13]).

In this paper the miscible displacement of one incompressible fluid by another in a porous medium is further investigated. A time-dependent, three-dimensional displacement problem with various boundary conditions and gravity effects, including mixed nonhomogeneous ones, is shown to possess a weak solution. The viscosity can be concentration-dependent, and the assumptions required on the data in the earlier papers are weakened; only physically reasonable assumptions are made. The analysis makes no use of the corresponding regularized problem; a weak solution is obtained as a limit of solutions to discrete time problems. It follows Alt and Luckhaus [3] for treating quasilinear elliptic-parabolic differential equations. The technique was later exploited by Alt and di Benedetto [2] and Arbogast [6] for handling the two-phase immiscible flow problem. However, we point out that the miscible displacement problem is different from the two-phase immiscible problem due to the above-mentioned difficulties. In particular, in the present problem special care must be taken on the transport and diffusion-dispersion coefficients in the concentration equation. We introduce here a solution-dependent space to handle this difficulty.

The above two cases deal with single-phase flow. Two-phase flow is more complex and is of greater practical interest. This case corresponds to the so-called secondary recovery in petroleum reservoirs where two fluid phases (usually water and oil) flow simultaneously, or to the fluid movement in an air-water porous media system in groundwater hydrology. In the last two decades, a considerable amount of effort has been made solely in the analysis of flow equations of two-phase incompressible, immiscible type. Transport equations have not been handled for the two-phase system. Existence of weak solutions for the flow equations has been established under various assumptions on physical data (see, e.g., [5, 16, 8, 15, 2, 6]). Numerical analyses of the flow equations of compressible type using finite elements have been carried out in [9, 10].

The second part of this paper extends the analysis for the single-phase, miscible displacement of one fluid by another to a two-phase flow and transport system. In addition to a strong coupling between flow and transport equations, the whole differential system combines the above difficulties for the displacement problem with those for the flow equations. In particular, the flow equations are usually degenerate, the number of equations is not known a priori at a given place of a porous medium, and the capillary pressure function is generally unbounded. Here we make an initial attempt to analyze both the flow and the transport equations for a two-phase system using the techniques for the single-phase, miscible displacement problem.

In the next section we consider the single-phase displacement problem. We begin with what is meant by a weak solution. Then we carefully state the assumptions on the physical data required for the major result obtained in this part. Most of this section is devoted to the proof of the major result. In section 3, we extend the results to a two-phase flow system. We shall follow the usual practice [4, 8] to write the flow equations of this system in a fractional flow formulation, i.e., in terms of a saturation and a global pressure so that the elliptic part of the system for this global pressure and the parabolic part for the saturation are separated. The concentration equation is obtained from the usual conservation law.

We end with a remark that uniqueness of the weak solution remains open. This is due to the coupling between the partial differential equations under consideration, which makes it difficult to obtain enough regularity of the solution. When the solution is assumed to have enough regularity (e.g., in the semiclassical sense), the uniqueness can be shown in the usual way [17].

2. Miscible displacement of one fluid by another. In section 2.1 the differential system for the single-phase, miscible displacement of one incompressible fluid by another in a porous medium $\Omega \subset \Re^d$ $(d \leq 3)$ is described. Then in section 2.2 we state the assumptions on the physical data, define what is meant by a weak solution, and state the major result shown in this section. The proof of the major result is presented in section 2.3, and two of the lemmas needed for the major result are proven in section 2.4.

2.1. The differential system. The usual equations describing two-component, incompressible, miscible displacement are given by (see, e.g., [7, 19])

$$\begin{aligned} &-\nabla \cdot \{k(x)(\nabla p - \rho g)/\mu(c)\} = q^I - q^P, \\ &\phi(x)\partial_t c - \nabla \cdot \big(D(u)\nabla c\big) + u \cdot \nabla c + q^I c = \hat{c} q^I \end{aligned} \tag{2.1}$$

for $(x,t) \in \Omega_T \equiv \Omega \times J$ with $J = (0,T]$ $(T > 0)$, where ϕ and k are the porosity and absolute permeability of the porous medium, μ and ρ are the viscosity and density of the fluid mixture, g denotes the gravitational, downward-pointing, constant vector, c indicates the concentration of one of the two components, p is the pressure of the fluid, D is the diffusion-dispersion coefficient, $\hat{c}$ is the injected concentration, q^I and q^P represent the sum of injection well source terms and production well sink terms, respectively, and u is the Darcy velocity of the fluid defined by

$$u = -\frac{k(x)}{\mu(c)}(\nabla p - \rho g). \tag{2.2}$$

For $\Gamma = \partial\Omega$, let

$$\Gamma = \Gamma_1^p \cup \Gamma_2^p = \Gamma_1^c \cup \Gamma_2^c, \quad \Gamma_1^p \cap \Gamma_2^p = \Gamma_1^c \cap \Gamma_2^c = \emptyset.$$

With this division of Γ, the boundary conditions are specified by

$$\begin{aligned} &u\cdot\nu - a_1(c)p = \varphi_1(c), && (x,t)\in\Gamma_1^p\times J,\\ &p = \varphi_2(x,t), && (x,t)\in\Gamma_2^p\times J,\\ &-(D\nabla c)\cdot\nu - a_2(c)c = \varphi_3(c), && (x,t)\in\Gamma_1^c\times J,\\ &c = \varphi_4(x,t), && (x,t)\in\Gamma_2^c\times J, \end{aligned} \tag{2.3}$$

where the a_i and φ_j are given functions ($i = 1, 2$, $1 \le j \le 4$) and ν is the outward unit normal to Γ. The initial condition is given by

$$c(x,0) = c_0(x), \quad x\in\Omega. \tag{2.4}$$

The differential system given by (2.1)–(2.4) for the main unknowns p and c will be studied in this section.

2.2. Assumptions and the major result. The usual Sobolev spaces $W^{l,\pi}(\Omega)$ with the norm $\|\cdot\|_{W^{l,\pi}(\Omega)}$ [1] will be used, where l is a nonnegative integer and $0 \le \pi \le \infty$. When $\pi = 2$, we simply write $H^l(\Omega) = W^{l,2}(\Omega)$. When $l = 0$, we have $L^2(\Omega) = H^0(\Omega)$. Below $(\cdot,\cdot)_Q$ denotes the $L^2(Q)$ inner product. (Q is omitted if $Q = \Omega$.) We now make the following assumptions:

(A1) $\Omega \subset \Re^d$ is a multiply connected domain with Lipschitz boundary Γ, $\Gamma = \Gamma_1^p \cup \Gamma_2^p = \Gamma_1^c \cup \Gamma_2^c$, $\Gamma_1^p \cap \Gamma_2^p = \Gamma_1^c \cap \Gamma_2^c = \emptyset$, each Γ_i^p and Γ_i^c is a $(d-1)$-dimensional domain, and $\Gamma_2^p \subset \Gamma_2^c$.

(A2) $\phi \in L^\infty(\Omega)$, $\phi(x) \ge \phi_* > 0$, and $k(x)$ is a bounded, symmetric, and uniformly positive definite matrix, i.e.,

$$0 < k_* \le |\xi|^{-2}\sum_{i,j=1}^{d} k_{ij}(x)\xi_i\xi_j \le k^* < \infty, \quad x\in\Omega,\ \xi\ne 0\in\Re^d.$$

(A3) The diffusion-dispersion term is given by

$$D(u) = \phi\{d_{mo}I + |u|\big(d_l E(u) + d_t E^\perp(u)\big)\},$$

where I is the d-by-d identity matrix, $d_{mo} > 0$ is the molecular diffusion coefficient, d_l and d_t are the longitudinal and transverse dispersion coefficients, respectively, the matrix $E(u)$ is the projection along the direction of flow determined by

$$E(u) = \left(\frac{u_i u_j}{|u|^2}\right), \quad |u| = \sqrt{u_1^2 + \ldots + u_d^2}, \quad u = (u_1, \ldots, u_d),$$

and $E^\perp(u) = I - E(u)$.

(A4) The following form is widely used for the viscosity μ:

$$\mu(c) = \mu(0)\big(1 + (\mathcal{M}^{1/4} - 1)c\big)^{-4} \quad \text{for } c\in[0,1],$$

where $\mathcal{M} = \mu(0)/\mu(1)$ is the mobility ratio.

(A5) q^I, $q^P \ge 0$, $q^I \in L^\infty(J; L^2(\Omega))$, and $q^P \in L^\infty(J; H^{-1}(\Omega))$.

(A6) In the case of $\Gamma_2^p = \emptyset$ and $a_1 \equiv 0$, φ_1 is independent of c and satisfies

$$\int_{\Gamma_1^p}\varphi_1 \mathrm{d}\sigma = \int_\Omega (q^I - q^P)\mathrm{d}x.$$

(A7) There is a subset $\Gamma_{1,*}^p \subset \Gamma_1^p$ (with nonzero measure only if $\Gamma_2^p = \emptyset$ and $a_1 \not\equiv 0$) such that $a_1 \geq a_{1,*} > 0$ on $\Gamma_{1,*}^p \times J \times [0,1]$.

(A8) The boundary data satisfy that φ_1 and φ_3 are continuous in c and

$$
\begin{aligned}
&|||\varphi_1|||_{L^\infty(J;H^{-1/2}(\Gamma_1^p))} < \infty, \quad |||\varphi_3|||_{L^\infty(J;H^{-1/2}(\Gamma_1^c))} < \infty,\\
&\varphi_2 \in L^\infty(J;H^1(\Omega)), \quad \varphi_4 \in L^2(J;W^{1,4}(\Omega)),\\
&\partial_t\varphi_4 \in L^1(\Omega_T), \quad 0 \leq \varphi_4(x,t) \leq 1 \quad \text{almost everywhere (a.e.) on } \Omega_T,\\
&\varphi_1(0) \geq 0, \quad \varphi_1(1) \geq 0 \quad \text{on } \Gamma_1^c,\\
&\varphi_3(0) \leq 0, \quad \varphi_3(1) \geq 0 \quad \text{on } \Gamma_1^c,
\end{aligned}
$$

where

$$
|||v||| = \Big\| \sup_{c\in[0,1]} |v(x,c)| \Big\|,
$$

for any given norm.

(A9) a_1, $a_2 \geq 0$; they are continuous in c; $|||a_1|||_{L^\infty(\Omega_T)}$ and $|||a_2|||_{L^\infty(\Omega_T)}$ are bounded; and

$$
a_1(0) = a_1(1) = 0 \quad \text{on } \Gamma_1^c.
$$

(A10) $\hat{c}$ and c_0 satisfy $0 \leq \hat{c} \leq 1$ a.e. on Ω_T and $0 \leq c_0 \leq 1$ a.e. on Ω.

We introduce the spaces

$$
\begin{aligned}
V &= \left\{ v \in H^1(\Omega) : v|_{\Gamma_2^p} = 0; \text{ if } \Gamma_2^p = \emptyset \text{ and } a_1 \equiv 0, \text{ then } \int_\Omega v \mathrm{d}x = 0 \right\},\\
W &= \{ v \in H^1(\Omega) : v|_{\Gamma_2^c} = 0 \}.
\end{aligned}
$$

Below V^* and W^* indicate the duals of V and W, respectively.

DEFINITION 2.1. *A weak solution of the system in* (2.1)–(2.4) *is a pair of functions* (p,c) *with* $p \in L^\infty(J;V) + \varphi_2$, $c \in L^2(J;W(u)) + \varphi_4$ *such that*

$$
\phi\partial_t c \in L^2(J;W^*(u)), \tag{2.5}
$$

$$
0 \leq c(x,t) \leq 1 \quad \text{a.e. on } \Omega_T, \tag{2.6}
$$

$$
\begin{aligned}
&(a(c)\{\nabla p - \rho g\}, \nabla v) + (a_1(c)p, v)_{\Gamma_1^p}\\
&\qquad = (q^I - q^P, v) - (\varphi_1(c), v)_{\Gamma_1^p} \quad \forall v \in L^\infty(J;V),
\end{aligned} \tag{2.7}
$$

$$
\begin{aligned}
&\int_J \langle \phi\partial_t c, v\rangle \, dt + \int_J (D(u)\nabla c, \nabla v)\, dt + \int_J (u\cdot\nabla c, v)\, dt\\
&\qquad + \int_J (q^I c, v)\, dt + \int_J (a_2(c)c, v)_{\Gamma_1^c}\, dt\\
&\qquad\qquad = \int_J (\hat{c}q^I, v)\, dt - \int_J (\varphi_3(c), v)_{\Gamma_1^c}\, dt \quad \forall v \in L^2(J;W(u)),
\end{aligned} \tag{2.8}
$$

$$
\begin{aligned}
&\int_J \langle \phi\partial_t c, v\rangle \, dt + \int_J (\phi(c - c_0), \partial_t v)\, dt = 0\\
&\qquad \forall v \in L^2(J;W(u)) \cap W^{1,1}(J;L^1(\Omega)), \quad v(x,T) = 0,
\end{aligned} \tag{2.9}
$$

where $a(c) = k(x)/\mu(c)$, u is given as in (2.2)*, and the space $W(u)$ is defined by*

$$W(u) = \{v \in W : (D(u)\nabla v, \nabla v) < \infty\}.$$

We now state the major result obtained in this section.

THEOREM 2.2. *Under assumptions* (A1)–(A10)*, the system in* (2.1)–(2.4) *has a weak solution in the sense of Definition* 2.1.

2.3. Proof of the major result. In this subsection we shall prove Theorem 2.2. We first state the following trivial lemma.

LEMMA 2.3. *It holds that*

$$d_{mo} + \min(d_l, d_t)|u| \le \phi^{-1}|\xi|^{-2} \sum_{i,j=1}^{d} D_{ij}(u)\xi_i\xi_j \le d_{mo} + \max(d_l, d_t)|u|, \quad \xi \ne 0 \in \Re^d.$$

For each positive integer M, divide J into $m = 2^M$ subintervals of equal length $h = T/m = 2^{-M}T$. Set $t_i = ih$ and $J_i = (t_{i-1}, t_i]$ for an integer i, $1 \le i \le m$. Denote the time difference operator by

$$\partial^\eta v(t) = \frac{v(t+\eta) - v(t)}{\eta}$$

for any function $v(t)$ and constant $\eta \in \Re$. Also, for any Hilbert space $\mathcal{H}$, define

$$l_h(\mathcal{H}) = \{v \in L^\infty(J; \mathcal{H}) : v \text{ is constant in time on each subinterval } J_i \subset J\}.$$

For $v^h \in l_h(\mathcal{H})$, set $v^i \equiv (v^h)^i = v^h|_{J_i}$ for notational convenience, when there is no ambiguity (i.e., h is omitted). Finally, let

$$\varphi_j^h(x,t) = \frac{1}{h}\int_{J_i} \varphi_j(x,\tau)\mathrm{d}\tau, \quad t \in J_i,\ j = 2, 4;$$

similar definitions $q^{I,h}$, $q^{P,h}$, and $\hat{c}^h$ are used for q^I, q^P, and $\hat{c}$, respectively.

Now, the discrete time solution is a pair of functions $p^h \in l_h(V) + \varphi_2^h$, $c^h \in l_h(W(u^h)) + \varphi_4^h$ satisfying

$$\begin{aligned}(a(c^h)\{\nabla p^h - \rho g\}, \nabla v) + (a_1(c^h)p^h, v)_{\Gamma_1^p} \\ = (q^{I,h} - q^{P,h}, v) - (\varphi_1(c^h), v)_{\Gamma_1^p} \quad \forall v \in l_h(V),\end{aligned} \tag{2.10}$$

and

$$\begin{aligned}\int_J (\phi\partial^{-h}c^h, v)\mathrm{d}t + \int_J (D(u^h)\nabla c^h, \nabla v)\mathrm{d}t + \int_J (u^h \cdot \nabla c^h, v)\mathrm{d}t \\ + \int_J (q^{I,h}c^h, v)\mathrm{d}t + \int_J (a_2(c^h)c^h, v)_{\Gamma_1^c}\mathrm{d}t \\ = \int_J (\hat{c}^h q^{I,h}, v)\mathrm{d}t - \int_J (\varphi_3(c^h), v)_{\Gamma_1^c}\mathrm{d}t \quad \forall v \in l_h(W(u^h)),\end{aligned} \tag{2.11}$$

with

$$u^h = -a(c^h)(\nabla p^h - \rho g). \tag{2.12}$$

This approximate scheme is extended such that $c^h = c_0$ for $t < 0$.

Below, C (with or without a subscript) indicates a generic constant independent of h, which will probably take on different values in different occurrences.

LEMMA 2.4. *For $h > 0$ small enough, the discrete scheme has a solution such that*

$$0 \le c^h(x,t) \le 1 \quad a.e.\ on \quad \Omega_T. \tag{2.13}$$

The proof of this lemma will be given in the next subsection.

LEMMA 2.5. *The solution to the discrete scheme also satisfies*

$$\|p^h\|_{L^\infty(J;H^1(\Omega))} + \|c^h\|_{L^2(J;H^1(\Omega))} + \|D^{1/2}(u^h)\nabla c^h\|_{L^2(\Omega_T)} \le C, \tag{2.14}$$

with constant C independent of h.

Proof. Take $v = p^h - \varphi_2^h \in l_h(V)$ in (2.10) to have

$$\begin{aligned}\|\nabla p^h\|^2_{L^2(\Omega)} + (a_1(c^h)p^h, p^h)_{\Gamma_1^p} \le C\{&\|\rho g\|^2_{L^2(\Omega)} + \|q^{I,h} - q^{P,h}\|^2_{H^{-1}(\Omega)} \\ &+|||\varphi_1|||^2_{H^{-1/2}(\Gamma_1^p)} + \|\varphi_2^h\|^2_{H^1(\Omega)}\} + \epsilon\|p^h\|^2_{L^2(\Omega)}, \quad t \in J,\end{aligned} \tag{2.15}$$

for any $\epsilon > 0$ (here and below $\epsilon > 0$ is an arbitrary constant as small as we please). Apply a variant of the Poincaré inequality

$$\|p^h\|_{L^2(\Omega)} \le C\{\|\nabla p^h\|_{L^2(\Omega)} + \|p_h\|_{L^2(\Gamma^p_{1,*})} + \|\varphi_2^h\|_{H^1(\Omega)}\} \tag{2.16}$$

and the inequality

$$\|\varphi_2^h\|_{H^1(\Omega)} \le \|\varphi_2\|_{H^1(\Omega)}$$

to obtain the bound for p^h in (2.14).

Now, choose $v = c^h - \varphi_4^h \in l_h(W(u^h))$ in (2.11) to see that

$$\begin{aligned}&\int_J (\phi\partial^{-h}c^h, c^h - \varphi_4^h)\mathrm{d}t + \int_J (D(u^h)\nabla c^h, \nabla(c^h - \varphi_4^h))\mathrm{d}t \\ &\quad + \int_J (u^h\cdot\nabla c^h, c^h - \varphi_4^h)\mathrm{d}t + \int_J (q^{I,h}c^h, c^h - \varphi_4^h)\mathrm{d}t \\ &\quad + \int_J (a_2(c^h)c^h, c^h - \varphi_4^h)_{\Gamma_1^c}\mathrm{d}t = \int_J (\hat{c}^h q^{I,h}, c^h - \varphi_4^h)\mathrm{d}t \\ &\qquad\qquad - \int_J (\varphi_3(c^h), c^h - \varphi_4^h)_{\Gamma_1^c}\mathrm{d}t.\end{aligned} \tag{2.17}$$

We now estimate each term in (2.17). We focus only on the transport and diffusion-dispersion terms; other terms can be easily estimated. By the definition of u^h, Lemma 2.3, the Hölder inequality, and the above bound on p^h, note that

$$\begin{aligned}\big|(D(u^h)\nabla c^h, \nabla\varphi_4^h)\big| &\le \epsilon(D(u^h)\nabla c^h, \nabla c^h) \\ &\quad + C\left(\|\nabla\varphi_4^h\|^2_{L^2(\Omega)} + \|u^h\|_{L^2(\Omega)}\|\nabla\varphi_4^h\|^2_{L^4(\Omega)}\right) \\ &\le \epsilon(D(u^h)\nabla c^h, \nabla c^h) + C\|\nabla\varphi_4^h\|^2_{L^4(\Omega)}.\end{aligned} \tag{2.18a}$$

Using (2.13), the same reasoning also yields that

$$\begin{aligned}\big|(u^h\cdot\nabla c^h, c^h)\big| &\le \epsilon(|u^h|\nabla c^h, \nabla c^h) + C(|u^h|c^h, c^h) \\ &\le \epsilon(D(u^h)\nabla c^h, \nabla c^h) + C(\Omega)\end{aligned} \tag{2.18b}$$

and that

$$\begin{aligned}\left|(u^h\cdot\nabla c^h,\varphi_4^h)\right|&\le \epsilon(|u^h|\nabla c^h,\nabla c^h)+C(|u^h|\varphi_4^h,\varphi_4^h)\\&\le \epsilon(D(u^h)\nabla c^h,\nabla c^h)+C\|\varphi_4^h\|^2_{L^4(\Omega)}.\end{aligned}\tag{2.18c}$$

Apply these estimates in (2.18a–c) to (2.17) to obtain

$$\begin{aligned}&\int_J(\phi\partial^{-h}c^h,c^h-\varphi_4^h)\mathrm{d}t+C_1\int_J(D(u^h)\nabla c^h,\nabla c^h)\mathrm{d}t\\&\le C(T,\Omega)\left\{1+\int_J\left(||q^{I,h}||^2_{H^{-1}(\Omega)}+|||\varphi_3|||^2_{H^{-1/2}(\Gamma_1^c)}\right.\right.\\&\qquad\left.\left.+\|\varphi_4^h\|^2_{H^1(\Omega)}+\|\varphi_4^h\|^4_{W^{1,4}(\Omega)}\right)\mathrm{d}t\right\}.\end{aligned}\tag{2.19}$$

Next, it is easy to see that

$$\int_J(\phi\partial^{-h}c^h,c^h)\mathrm{d}t=\sum_{i=1}^m(\phi(c^i-c^{i-1}),c^i)\ge\frac{1}{2}\{(\phi c^m,c^m)-(\phi c^0,c^0)\}.\tag{2.20}$$

Also, we find that [3, 6]

$$\begin{aligned}\int_J(\phi\partial^{-h}c^h,\varphi_4^h)\mathrm{d}t&=(\phi c^m,\varphi_4^m)-(\phi c^0,\varphi_4^1)-\int_0^{T-h}(\phi c^h,\partial^h\varphi_4^h)\mathrm{d}t\\&\le C\left\{\|\varphi_4^h\|_{L^\infty(J;L^1(\Omega))}+\int_0^{T-h}\|\partial^h\varphi_4^h\|_{L^1(\Omega)}\mathrm{d}t\right\}\\&\le C\left\{\|\varphi_4^h\|_{L^\infty(J;L^1(\Omega))}+\int_J\|\partial_t\varphi_4\|_{L^1(\Omega)}\mathrm{d}t\right\}.\end{aligned}\tag{2.21}$$

Finally, combine (2.17)–(2.21) to have the desired result for c^h. □

COROLLARY 2.6. *For any $2\le r<\infty$, for a subsequence $p^h\rightharpoonup p$ weakly in $L^r(J;H^1(\Omega))$ and $c^h\rightharpoonup c$ weakly in $L^2(J;H^1(\Omega))$. Furthermore, $p\in L^\infty(J;V)+\varphi_2$, $c\in L^2(J;W)+\varphi_4$, and*

$$0\le c(x,t)\le 1\quad a.e.\ on\quad \Omega_T.\tag{2.22}$$

Proof. It follows from Lemma 2.5 that $c^h-\varphi_4^h$ converges weakly in $L^2(J;W)$. Since $\varphi_4^h\rightharpoonup\varphi_4$ weakly in $L^2(J;H^1(\Omega))$, $c^h\rightharpoonup c$ weakly in $L^2(J;H^1(\Omega))$ with $c\in L^2(J;W)+\varphi_4$. The same argument shows that $p^h\rightharpoonup p$ weakly in $L^r(J;H^1(\Omega))$ with $p\in L^r(J;V)+\varphi_2$ for $2\le r<\infty$. Since $\|p\|_{L^r(J;H^1(\Omega))}\le C$ with C independent of r, in fact $p\in L^\infty(J;V)+\varphi_2$. Finally, (2.22) follows from (2.13). □

LEMMA 2.7. *There is a subsequence such that $c^h\to c$ strongly in $L^2(\Omega_T)$.*

This lemma also will be shown in section 2.4.

COROLLARY 2.8. *There is a subsequence such that $c^h\to c$ strongly in $L^2(J;H^{1-\pi}(\Omega))$ and $L^2(J;H^{1/2-\pi}(\partial\Omega))$ for any $0<\pi<1/2$, and $c^h\to c$ pointwise a.e. on Ω_T.*

Proof. Apply the interpolation inequality

$$\|v\|_{H^\sigma(\Omega)}\le\delta\|v\|_{H^1(\Omega)}+C_\delta\|v\|_{L^2(\Omega)}\tag{2.23}$$

for any $0<\sigma<1$ and $\delta>0$, the boundedness of the trace operator, and Lemma 2.7 to prove the desired statement. □

We are now ready to prove Theorem 2.2.

Proof of Theorem 2.2. From Corollaries 2.6 and 2.8, (2.10) implies (2.7) since $\cup_{M=1}^{\infty} l_h(V)$ is dense in $L^\infty(J;V)$. Also, it follows from (2.11) that

$$\begin{aligned}\lim_{h\to 0^+}\Big\{&\int_J(\phi\partial^{-h}c^h,v)\mathrm{d}t+\int_J(D(u^h)\nabla c^h,\nabla v)\mathrm{d}t+\int_J(u^h\cdot\nabla c^h,v)\mathrm{d}t\\ &+\int_J(q^{I,h}c^h,v)\mathrm{d}t+\int_J(a_2(c^h)c^h,v)_{\Gamma_1^c}\mathrm{d}t\Big\}\\ =\lim_{h\to 0^+}&\Big\{\int_J(\hat{c}^h q^{I,h},v)\mathrm{d}t-\int_J(\varphi_3(c^h),v)_{\Gamma_1^c}\mathrm{d}t\Big\}\ \forall v\in\cup_{M=1}^{\infty}l_h(W(u^h)).\end{aligned}\tag{2.24}$$

By (2.7) and (2.10), we see that

$$\begin{aligned}&(a(c^h)\nabla[p^h-p],\nabla[p^h-p])+(a(c^h)\nabla[p^h-p],\nabla p)\\ &-([a(c)-a(c^h)](\nabla p-\rho g),\nabla p^h)-(a_1(c^h)(p-p^h),p^h)_{\Gamma_1^p}\\ &-([a_1(c)-a_1(c^h)]p,p^h)_{\Gamma_1^p}=-([q^I-q^P]-[q^{I,h}-q^{P,h}],p^h)\\ &\qquad+(\varphi_1(c)-\varphi_1(c^h),p^h)_{\Gamma_1^p}.\end{aligned}$$

Then, by Corollaries 2.6 and 2.8 and Lebesgue's dominated convergence theorem, we see that $\nabla p^h\to\nabla p$ strongly in $(L^2(\Omega_T))^d$. Hence, from the definition of u^h, $u^h\to u$ strongly in $(L^2(\Omega_T))^d$ and by Lemma 2.3, $D(u^h)\to D(u)$ strongly in $(L^2(\Omega_T))^{d\times d}$. Therefore, by the strong convergence of $\{D(u^h)\}$ and weak convergence of $\{\nabla c^h\}$, we have

$$\lim_{h\to 0^+}\int_J(D(u^h)\nabla c^h,\nabla v)\mathrm{d}t=\int_J(D(u)\nabla c,\nabla v)\mathrm{d}t\quad\forall v\in\cup_{M=1}^{\infty}l_h(W(u^h)).$$

Similarly, by the strong convergence of $u^h\to u$ and weak convergence of $\nabla c^h\rightharpoonup\nabla c$, we see that

$$\lim_{h\to 0^+}\int_J(u^h\cdot\nabla c^h,v)\mathrm{d}t=\int_J(u\cdot\nabla c,v)\mathrm{d}t\quad\forall v\in\cup_{M=1}^{\infty}l_h(W(u^h)).$$

Next, by the definition of $q^{I,h}$ and $\hat{c}^h$, and the continuity of φ_3 and a_2 in c, it follows from Corollary 2.8 that

$$\lim_{h\to 0^+}\Big\{\int_J(q^{I,h}c^h,v)\mathrm{d}t+\int_J(a_2(c^h)c^h,v)_{\Gamma_1^c}\mathrm{d}t\Big\}=\int_J(q^Ic,v)\mathrm{d}t+\int_J(a_2(c)c,v)_{\Gamma_1^c}\mathrm{d}t\quad\forall v\in\cup_{M=1}^{\infty}l_h(W(u^h)),$$

and

$$\lim_{h\to 0^+}\Big\{\int_J(\hat{c}^h q^{I,h},v)\mathrm{d}t-\int_J(\varphi_3(c^h),v)_{\Gamma_1^c}\mathrm{d}t\Big\}=\int_J(\hat{c}q^I,v)\mathrm{d}t-\int_J(\varphi_3(c),v)_{\Gamma_1^c}\mathrm{d}t\quad\forall v\in\cup_{M=1}^{\infty}l_h(W(u^h)).$$

Also, for any $v\in L^2(J;W(u))$, $v^h\in l_h(W(u^h))$ for h sufficiently small, where $v^h(x,t)=h^{-1}\int_{J_k}v(x,\tau)\mathrm{d}\tau$. Then, by (2.11), (2.14), and the compact embedding relation $H^1(\Omega)\hookrightarrow L^4(\Omega)$, we observe that

$$\int_J(\phi\partial^{-h}c^h,v)\mathrm{d}t=\int_J(\phi\partial^{-h}c^h,v^h)\mathrm{d}t\le C\big\{\|D^{1/2}(u)\nabla v\|_{L^2(\Omega_T)}+\|v\|_{L^2(\Omega_T)}\big\}$$

for h sufficiently small. Consequently, for a subsequence $\phi\partial^{-h}c^h$ converges weakly in $L^2(J;W^*(u))$. If $v \in C_0^\infty(\Omega_T)$, with $h > 0$ small enough we see that

$$\int_J (\phi\partial^{-h}c^h, v)\mathrm{d}t = -\int_0^{T-h} (\phi c^h, \partial^h v)\mathrm{d}t \to -\int_J (\phi c, \partial_t v)\mathrm{d}t = \int_J \langle \phi\partial_t c, v\rangle \,\mathrm{d}t$$

as a distribution. Therefore, $\phi\partial^{-h}c^h \rightharpoonup \phi\partial_t c$ weakly in $L^2(J;W^*(u))$. Combining all these results, (2.8) follows from (2.24) since $\cup_{M=1}^\infty l_h(W(u^h))$ is dense in $L^2(J;W(u))$. Also, as for (2.19), it can be shown that $c \in L^2(J;W(u))$ from (2.8).

Finally, if $v \in L^2(J;W(u)) \cap W^{1,1}(J;L^1(\Omega))$ with $v(x,T) = 0$, we find that

$$\int_J (\phi\partial^{-h}c^h, v)\mathrm{d}t + \int_0^{T-h} (\phi[c^h - c^0], \partial^h v)\mathrm{d}t = \frac{1}{h}\int_{T-h}^T (\phi[c^h - c^0], v)\mathrm{d}t,$$

which yields (2.9). Thus the proof of Theorem 2.2 is complete. □

2.4. Proof of Lemmas 2.4 and 2.7. In this subsection the possibility that c is outside $[0,1]$ is allowed. All functions of c are extended constantly outside $[0,1]$.

Lemma 2.4 is purely an elliptic result and will obviously follow from the next proposition. For notational convenience the superscript h is omitted below.

PROPOSITION 2.9. *In addition to assumptions* (A1)–(A10)*, suppose that* $0 < \eta_* \le \eta_1(x) \in L^\infty(\Omega)$ *and* $0 \le \eta_2(x) \le \eta_1(x)$*. Then, for* η_* *sufficiently big, the following problem has a weak solution* $(p,c) \in (V+\varphi_2)\times(W(u)+\varphi_4)$*:*

$$\begin{aligned}(a(c)\{\nabla p - \rho g\}, \nabla v) + (a_1(c)p, v)_{\Gamma_1^p} \\ = (q^I - q^P, v) - (\varphi_1(c), v)_{\Gamma_1^p} \quad \forall v \in V,\end{aligned} \tag{2.25}$$

$$\begin{aligned}(\eta_1 c, v) + (D(u)\nabla c, \nabla v) + (u\cdot\nabla c, v) + (q^I c, v) + (a_2(c)c, v)_{\Gamma_1^c} \\ = (\hat{c}q^I, v) - (\varphi_3(c), v)_{\Gamma_1^c} + (\eta_2, v) \quad \forall v \in W(u),\end{aligned} \tag{2.26}$$

and

$$0 \le c(x,t) \le 1 \quad \text{a.e. on } \Omega_T, \tag{2.27}$$

where u *is given as in* (2.12).

Proof. Let $\{v_i^1\}_{i=1}^\infty$ and $\{v_i^2\}_{i=1}^\infty$ be bases for V and W, respectively, and set $V_m = \mathrm{span}\{v_1^1, \dots, v_m^1\}$ and $W_m = \mathrm{span}\{v_1^2, \dots, v_m^2\}$. With V_m and W_m replacing V and W in (2.25) and (2.26), respectively, we obtain a Galerkin procedure.

For $v^j = \sum_{i=1}^m \beta_i^j v_i^j$, $j = 1, 2$, we introduce the mapping $\Phi_m : \Re^{2m} \to \Re^{2m}$ by

$$\Phi_m \begin{pmatrix} \beta^1 \\ \beta^2 \end{pmatrix} = \begin{pmatrix} \hat{\beta}^1 \\ \hat{\beta}^2 \end{pmatrix},$$

where

$$\begin{aligned}\hat{\beta}_i^1 =& (a(v^2+\varphi_4)\{\nabla(v^1+\varphi_2) - \rho g\}, \nabla v_i^1) + (a_1(v^2+\varphi_4)(v^1+\varphi_2), v_i^1)_{\Gamma_1^p} \\ & -(q^I - q^P, v_i^1) + (\varphi_1(v^2+\varphi_4), v_i^1)_{\Gamma_1^p}, \\ \hat{\beta}_i^2 =& (\eta_1(v^2+\varphi_4), v_i^2) + (D(\hat{u})\nabla(v^2+\varphi_4), \nabla v_i^2) + (\hat{u}\cdot\nabla(v^2+\varphi_4), v_i^2) \\ & +(q^I(v^2+\varphi_4), v_i^2) + (a_2(v^2+\varphi_4)(v^2+\varphi_4), v_i^2)_{\Gamma_1^c} \\ & -(\hat{c}q^I, v_i^2) + (\varphi_3(v^2+\varphi_4), v_i^2)_{\Gamma_1^c} - (\eta_2, v_i^2),\end{aligned}$$

with $u = -a(v^2+\varphi_4)\{\nabla(v^1+\varphi_2)-\rho g\}$ and $\hat{u} = mu/(m+|u|)$. Note that to handle the difficulty associated with the transport and diffusion-dispersion terms, we have introduced $\hat{u}$ above. By the assumptions (A1)–(A10), Φ_m is continuous. Also, it can be easily seen that

$$\begin{aligned}\Phi_m\begin{pmatrix}\beta^1\\ \beta^2\end{pmatrix}\cdot\begin{pmatrix}\beta^1\\ \beta^2\end{pmatrix} \geq & C_1(m)\big\{\|\nabla v^1\|^2_{L^2(\Omega)}+\|\nabla v^2\|^2_{L^2(\Omega)}\big\}-\epsilon\|v^1\|^2_{L^2(\Omega)}\\ & +(\eta_1(v^2+\varphi_4)-\eta_2,v^2)+(q^I\varphi_4,v^2)\\ & -C\Big\{|||\varphi_1|||^2_{H^{-1/2}(\Gamma_1^p)}+|||\varphi_3|||^2_{H^{-1/2}(\Gamma_1^c)}+\|\varphi_2\|^2_{H^1(\Omega)}\\ & \quad +\|\varphi_4\|^2_{H^1(\Omega)}+\|q^I\|^2_{H^{-1}(\Omega)}+\|q^P\|^2_{H^{-1}(\Omega)}\\ & \quad +\|\rho g\|^2_{L^2(\Omega)}+\|v^2\|^2_{L^2(\Omega)}\Big\}\end{aligned}$$

for any $\epsilon>0$. Note that

$$(\eta_1(v^2+\varphi_4)-\eta_2,v^2)\geq\frac{1}{2}\eta_*\|v^2\|^2_{L^2(\Omega)}-C\big\{1+\|\varphi_4\|^2_{L^2(\Omega)}\big\}, \tag{2.28}$$

and, by the compact embedding relation $H^1(\Omega)\hookrightarrow L^4(\Omega)$ again,

$$\begin{aligned}(q^I\varphi_4,v^2)\leq & \|q^I\|_{L^2(\Omega)}\|\varphi_4\|_{L^4(\Omega)}\|v^2\|_{L^4(\Omega)}\\ \leq & C\|q^I\|_{L^2(\Omega)}\|\varphi_4\|_{L^4(\Omega)}\|v^2\|_{H^1(\Omega)}.\end{aligned}$$

Now, with Poincaré's inequality and η_* big enough, combining these results yields that

$$\Phi_m\begin{pmatrix}\beta^1\\ \beta^2\end{pmatrix}\cdot\begin{pmatrix}\beta^1\\ \beta^2\end{pmatrix}\geq C_1(m)\big\{\|v^1\|^2_{H^1(\Omega)}+\|v^2\|^2_{H^1(\Omega)}\big\}-C,$$

which is strictly positive for $|\beta^1|+|\beta^2|$ sufficiently big. As a result, Φ_m has a zero; i.e., there is a solution to the Galerkin approximation with $\hat{u}$ replacing u for each m.

As in the proof of Lemma 2.5, it can be seen that the modified Galerkin solutions p^m and c^m are uniformly bounded in $H^1(\Omega)$ (independently of m), so for a subsequence $p^m\rightharpoonup p$ and $c^m\rightharpoonup c$ weakly in $H^1(\Omega)$ with $p\in V+\varphi_2$ and $c\in W(u)+\varphi_4$. Moreover, $c_m\to c$ strongly in $H^{1-\pi}(\Omega)$ $(0<\pi<1/2)$ and pointwise a.e. both on Ω and $\partial\Omega$. Therefore, (p,c) is a weak solution to the system in (2.25) and (2.26).

Finally, we apply a standard maximum principle argument to (2.26) to show (2.27). Take $v=c^-=\min(c,0)\in W(u)$ in (2.26) to have

$$\begin{aligned}(\eta_1 c-\eta_2,c^-)= & -(D(u)\nabla c,\nabla c^-)-(u\cdot\nabla c,c^-)-(q^Ic,c^-)\\ & -(a_2(c)c,c^-)_{\Gamma_1^c}+(\hat{c}q^I,c^-)-(\varphi_3(c),c^-)_{\Gamma_1^c}.\end{aligned}$$

Note that $c^-\in V$ by assumption (A1) (if $\Gamma_2^p=\emptyset$ and $a_1\equiv 0$, consider $c^- - \int_\Omega c^-\mathrm{d}x$). Then, by (2.25) with $v=(c^-)^2$, we see that

$$\begin{aligned}(u\cdot\nabla c,c^-)= & \tfrac{1}{2}(u,\nabla(c^-)^2)\\ = & \tfrac{1}{2}\big\{(a_1(c)p,(c^-)^2)_{\Gamma_1^p}-(q^I-q^P,(c^-)^2)+(\varphi_1(c),(c^-)^2)_{\Gamma_1^p}\big\}.\end{aligned}$$

Use these two equations to see that

$$\begin{aligned}(\eta_1 c-\eta_2,c^-)= & -(D(u)\nabla c,\nabla c^-)-\tfrac{1}{2}\big\{(a_1(c)p,(c^-)^2)_{\Gamma_1^p}+(q^Ic,c^-)+(q^Pc,c^-)\big\}\\ & -\tfrac{1}{2}(\varphi_1(c),(c^-)^2)_{\Gamma_1^c}-(a_2(c)c,c^-)_{\Gamma_1^c}+(\hat{c}q^I,c^-)-(\varphi_3(c),c^-)_{\Gamma_1^c}\\ \leq & 0\end{aligned}$$

by assumptions (A5), (A8), and (A9), which, as in (2.28), shows that $c^- = 0$ a.e. on Ω_T provided η_* is sufficiently large. This concludes that $c \geq 0$ a.e. on Ω_T. Similarly, with $v = (c-1)^+ = \max(c-1, 0) \in W(u)$ in (2.26), we can see that $c \leq 1$ a.e. on Ω_T. This completes the proof of the proposition. ∎

The next lemma is related to Lemma 3 in [6] and is needed for proving Lemma 2.7. Both Lemma 2.10 and the proof of Lemma 2.7 are based on the ideas presented in [3].

LEMMA 2.10. *Let c^h satisfy* (2.11). *Then there exists C such that, for any $\zeta > 0$,*

$$\frac{1}{\zeta}\int_{\zeta}^{T} \|\phi^{1/2}\big(c^h(\cdot,t) - c^h(\cdot,t-\zeta)\big)\|^2_{L^2(\Omega)}\,dt \leq C.$$

Proof. Let k be fixed ($1 \leq k \leq m$); for $\tau \in J_i$, we define the interval $Q = Q(\tau) = ((i-k)h, ih]$ and the characteristic function χ_Q. Take $v(x,t) = kh\chi_Q(t)\partial^{-kh}(c^h - \varphi_4^h)(x,\tau) \in l_h(W)$ in (2.11) and apply the relation

$$\int_J \partial^{-h} c^h \chi_Q \mathrm{d}t = \sum_{j=i-k+1}^{i} (c^j - c^{j-1}) = kh\partial^{-kh} c^h(\cdot,\tau),$$

(2.13), and (2.14) to obtain

$$\begin{aligned} kh\int_{kh}^{T} \|\phi^{1/2}\partial^{-kh}c^h(\cdot,\tau)\|^2_{L^2(\Omega)}\mathrm{d}\tau &\leq C + kh\int_{kh}^{T}(\phi\partial^{-kh}c^h(\cdot,\tau), \partial^{-kh}\varphi_4^h(\cdot,\tau))\mathrm{d}\tau \\ &\leq C\left(1 + \|\partial_t\varphi_4\|_{L^1(\Omega_T)}\right), \end{aligned}$$

which implies the desired result since c^h is constant on each interval J_i. ∎

For m_1 a positive integer, let $\delta = T/m_1$ and $J_k^\delta = ((k-1)\delta, k\delta]$. Introduce the operator $A^\delta : L^1(J) \to L^1(J)$ by

$$A^\delta(v) = \frac{1}{\delta}\int_{J_k^\delta} v(\tau)\mathrm{d}\tau, \quad t \in J_k^\delta.$$

We are now in a position to prove Lemma 2.7.

Proof of Lemma 2.7. For any ζ, $N > 0$, define

$$\begin{aligned} Q = Q(c^h, \zeta, N) = \big\{ t \in (\zeta, T] : &\|c^h(\cdot,t)\|^2_{H^1(\Omega)} + \|c^h(\cdot,t-\zeta)\|^2_{H^1(\Omega)} \\ &+ \tfrac{1}{\zeta}\|\phi^{1/2}\big(c^h(\cdot,t) - c^h(\cdot,t-\zeta)\big)\|^2_{L^2(\Omega)} > N \big\}. \end{aligned}$$

Obviously, by (2.13), (2.14), and Lemma 2.10, the measure of Q is less than C/N with constant C independent of h. If $t \in (\zeta, T] \setminus Q$, then

$$\|c^h(\cdot,t) - c^h(\cdot,t-\zeta)\|^2_{L^2(\Omega)} \leq \frac{\zeta N}{\phi_*}.$$

Thus we see that

$$\int_{\zeta}^{T} \|c^h(\cdot,t) - c^h(\cdot,t-\zeta)\|^2_{L^2(\Omega)}\mathrm{d}t \leq \frac{\zeta N T}{\phi_*} + \frac{4C|\Omega|}{N}$$

so that, by the arbitrariness of N,

$$\int_{\zeta}^{T} \|c^h(\cdot,t) - c^h(\cdot,t-\zeta)\|^2_{L^2(\Omega)}\mathrm{d}t \to 0 \qquad \text{as } \zeta \to 0^+$$

uniformly in h. Therefore, by the definition of A^δ we see that

$$\int_J \|c^h - A^\delta(c^h)\|^2_{L^2(\Omega)} \mathrm{d}t \leq \underbrace{\frac{2}{\delta}\int_0^\delta \int_\zeta^T \|c^h(\cdot,t) - c^h(\cdot,t-\zeta)\|^2_{L^2(\Omega)} \mathrm{d}t\mathrm{d}\zeta}_{\to 0 \quad \text{as } \delta \to 0^+} \tag{2.29}$$

uniformly in h. Also, $\|A^\delta(c^h)\|_{L^2(J,H^1(\Omega))}$ is uniformly bounded, so for fixed $\delta > 0$, $A^\delta(c^h)$ converges strongly in $L^2(\Omega_T)$ as $h \to 0^+$. Consequently, apply (2.29) and the inequality

$$\|c^{h_1} - c^{h_2}\|_{L^2(\Omega_T)} \leq \sum_{j=1}^{2} \|c^{h_j} - A^\delta(c^{h_j})\|_{L^2(\Omega_T)} + \|A^\delta(c^{h_1}) - A^\delta(c^{h_2})\|_{L^2(\Omega_T)}$$

to complete the proof of Lemma 2.7. □

3. Two-phase flow and transport. In this section we consider two-phase flow and transport equations in a porous medium $\Omega \subset \Re^d$ $(d \leq 3)$. These equations will be reviewed in section 3.1. The differential system we shall study in this section will be derived in section 3.2. The major result in this section is stated in section 3.3 and proven in section 3.4 and section 3.5.

3.1. Flow and transport equations. The mass balance equation for each of the fluid phases is given by [7]

$$\phi \frac{\partial(\rho_\alpha s_\alpha)}{\partial t} + \nabla \cdot (\rho_\alpha u_\alpha) = \rho_\alpha q_\alpha, \quad \alpha = w, n, \tag{3.1}$$

where $\alpha = w$ denotes the wetting phase (e.g., water), $\alpha = n$ indicates the nonwetting phase (e.g., oil or air), ϕ is the porosity of the porous medium, and ρ_α, s_α, u_α, and q_α are, respectively, the density, (reduced) saturation, volumetric velocity, and external volumetric flow rate of the α-phase. The volumetric velocity u_α is given again by the Darcy law

$$u_\alpha = -\frac{k k_{r\alpha}}{\mu_\alpha}(\nabla p_\alpha - \rho_\alpha g), \quad \alpha = w, n, \tag{3.2}$$

where k is the absolute permeability of the porous medium and p_α, μ_α, and $k_{r\alpha}$ are the pressure, viscosity, and relative permeability of the α-phase, respectively. In addition to (3.1) and (3.2), the customary property for the saturations is

$$s_w + s_n = 1, \tag{3.3}$$

and the two pressures are related by the capillary pressure function

$$p_c = p_n - p_w. \tag{3.4}$$

With the flow of the fluids specified as above, an equation for the transport of a chemical constituent is needed. The constituent can be transported in each of the phases, so the equation is written for each phase. Let c_α denote the mass concentration of the constituent in the α phase. Then the mass balance law for the constituent in each phase reads as follows:

$$\phi \frac{\partial(\rho_\alpha s_\alpha c_\alpha)}{\partial t} + \nabla \cdot (\rho_\alpha u_\alpha c_\alpha - \phi \rho_\alpha s_\alpha D_\alpha \nabla c_\alpha) + \phi r_\alpha \rho_\alpha s_\alpha c_\alpha = \hat{c}_\alpha \rho_\alpha \hat{q}_\alpha \tag{3.5}$$

for $\alpha = w$, n, where D_α, r_α, and $\hat{c}_\alpha$ are the diffusion-dispersion coefficient, reaction rate, and concentration in the external flow for the α phase, respectively.

3.2. The differential system. The following functional dependence is physically reasonable:

$$k_{r\alpha} = k_{r\alpha}(x, s_\alpha), \quad p_c = p_c(x, s_w), \quad D_\alpha = D_\alpha(x, u_\alpha).$$

However, to extend the analysis of the last section to the present problem we assume that the viscosity μ_α is independent of c_α. The case that $\mu_\alpha = \mu_\alpha(c_\alpha)$ needs to be handled with a different argument and will be treated in a forthcoming paper. The model derived here is of interest in itself.

In order to separate the pressure and saturation equations, we introduce the phase mobility functions

$$\lambda_\alpha(x, s_\alpha) = k_{r\alpha}/\mu_\alpha, \quad \alpha = w, n,$$

and the total mobility

$$\lambda(x, s) = \lambda_w + \lambda_n,$$

where $s = s_w$. The fractional flow functions are defined by

$$f_\alpha(x, s) = \lambda_\alpha/\lambda, \quad \alpha = w, n.$$

Following [4, 8], we define the global pressure

$$p = p_n - \int_0^s \left(f_w \frac{\partial p_c}{\partial s} \right)(x, \xi)\mathrm{d}\xi. \tag{3.6}$$

Also, we shall use the complementary pressure [6]

$$\theta = -\int_0^s \left(f_w f_n \frac{\partial p_c}{\partial s} \right)(x, \xi). \tag{3.7}$$

Finally, we define the total velocity

$$u = u_w + u_n. \tag{3.8}$$

Now, under the assumption that the fluids are incompressible we apply (3.3) and (3.8) to (3.1) to see that

$$\nabla \cdot u = q \equiv q_w + q_n, \tag{3.9}$$

and we apply (3.4), (3.6), and (3.8) to (3.2) to obtain

$$u = -k(\lambda \nabla p + \gamma_1), \tag{3.10}$$

where

$$\gamma_1 = -\lambda_w \nabla_x p_c + \lambda \int_0^s \nabla_x \left(f_w \frac{\partial p_c}{\partial s} \right)(x, \xi)\mathrm{d}\xi - \big(\lambda_w \rho_w + \lambda_n \rho_n\big) g.$$

Similarly, apply (3.4), (3.6), and (3.7) to (3.1) and (3.2) with $\alpha = w$ to have

$$\phi \frac{\partial s}{\partial t} - \nabla \cdot \big\{ k\big(\lambda \nabla \theta + \lambda_w \nabla p + \gamma_2\big) \big\} = q_w, \tag{3.11}$$

where

$$\gamma_2 = -\lambda_w \nabla_x p_c + \lambda_w \int_0^s \nabla_x \left(f_w \frac{\partial p_c}{\partial s} \right)(x,\xi)\mathrm{d}\xi + \lambda \int_0^s \nabla_x \left(f_w f_n \frac{\partial p_c}{\partial s} \right)(x,\xi)\mathrm{d}\xi - \lambda_w \rho_w g.$$

Finally, it can be seen that the phase velocities are determined by

$$\begin{aligned} u_w &= -k\big(\lambda \nabla\theta + \lambda_w \nabla p + \gamma_2\big), \\ u_n &= k\big(\lambda \nabla\theta - \lambda_n \nabla p + \gamma_3\big), \end{aligned} \tag{3.12}$$

where

$$\gamma_3 = -\lambda_n \int_0^s \nabla_x \left(f_w \frac{\partial p_c}{\partial s} \right)(x,\xi)\mathrm{d}\xi + \lambda \int_0^s \nabla_x \left(f_w f_n \frac{\partial p_c}{\partial s} \right)(x,\xi)\mathrm{d}\xi + \lambda_n \rho_n g.$$

The pressure equation is given by (3.9) and (3.10), while the saturation equation is described by (3.11). In hydrology, it is common to replace the pressures by the pressure heads

$$h_\alpha = p_\alpha/(\rho_{sw} g), \quad \alpha = w, n,$$

where ρ_{sw} is the density of water at the standard temperature and pressure. With the pressure heads h_α, similar equations to those in (3.9)–(3.12) can be obtained; in this paper we shall use the pressures.

To derive the concentration equation, we make use of the usual equilibrium assumption on mass transfer. That is, the constituent instantaneously establishes an equilibrium distribution between the two phases. Then the concentration in each phase is proportional to that in the other phase:

$$c_n = H c_w,$$

where H is called the Henry constant and taken to be one for simplicity. Then under the incompressibility assumption, apply (3.3) and (3.8) to (3.5) to find that

$$\phi \frac{\partial c}{\partial t} - \nabla \cdot \big(D \nabla c - u c \big) + R c = \hat{c}\hat{q}, \tag{3.13}$$

where $c = c_w$, $D = \phi(s_w D_w + s_n D_n)$, and $R = \phi(s_w r_w + s_n r_n)$.

In summary, from (3.9)–(3.11) and (3.13) we have the differential system

$$\begin{aligned} &-\nabla \cdot \big\{ k(\lambda(s,c)\nabla p + \gamma_1(s,c)) \big\} = q(s,c), \\ &\phi \partial_t s - \nabla \cdot \big\{ k\big(\lambda(s,c)\nabla\theta + \lambda_w(s,c)\nabla p + \gamma_2(s,c)\big) \big\} = q_w(s,c), \\ &\phi \partial_t c - \nabla \cdot \big\{ D(s,u_w,u_n)\nabla c - u c \big\} + R(s) c = \hat{c}\hat{q}(s,c), \end{aligned} \tag{3.14}$$

where

$$\begin{aligned} u &= -k(\lambda(s,c)\nabla p + \gamma_1(s,c)), \\ u_w &= -k\big(\lambda(s,c)\nabla\theta + \lambda_w(s,c)\nabla p + \gamma_2(s,c)\big), \\ u_n &= k\big(\lambda(s,c)\nabla\theta - \lambda_n(s,c)\nabla p + \gamma_3(s,c)\big). \end{aligned} \tag{3.15}$$

Finally, s is related to θ through (3.7):

$$s = \mathcal{S}(\theta), \tag{3.16}$$

where $\mathcal{S}(x,\theta)$ is the inverse of (3.7) for $0 \le \theta \le \theta^*(x)$ with

$$\theta^*(x) = -\int_0^1 \left(f_w f_n \frac{\partial p_c}{\partial s}\right)(x,\xi)\mathrm{d}\xi.$$

The differential system in (3.14)–(3.16) determines the main unknowns p, s, θ, and c. The model is completed by specifying boundary and initial conditions.

The division of Γ is

$$\Gamma = \Gamma_1^p \cup \Gamma_2^p = \Gamma_1^\theta \cup \Gamma_2^\theta = \Gamma_1^c \cup \Gamma_2^c,$$
$$\emptyset = \Gamma_1^p \cap \Gamma_2^p = \Gamma_1^\theta \cap \Gamma_2^\theta = \Gamma_1^c \cap \Gamma_2^c.$$

The boundary conditions are specified by

$$\begin{aligned}
&u\cdot\nu - a_1(s,c)p = \varphi_1(s,c), && (x,t)\in\Gamma_1^p\times J,\\
&p = \varphi_2(x,t), && (x,t)\in\Gamma_2^p\times J,\\
&u_w\cdot\nu - a_2(s,c)\theta = \varphi_3(s,c), && (x,t)\in\Gamma_1^\theta\times J,\\
&\theta = \varphi_4(x,t), && (x,t)\in\Gamma_2^\theta\times J,\\
&(uc - D\nabla c)\cdot\nu - a_3(s,c)c = \varphi_5(s,c), && (x,t)\in\Gamma_1^c\times J,\\
&c = \varphi_6(x,t), && (x,t)\in\Gamma_2^c\times J,
\end{aligned} \tag{3.17}$$

where the a_i and φ_i are given functions. The initial conditions are given by

$$\begin{aligned}
&\theta(x,0) = \theta_0(x), \quad x\in\Omega,\\
&c(x,0) = c_0(x), \quad x\in\Omega.
\end{aligned} \tag{3.18}$$

The differential system has a clear structure. Note that while λ_w and λ_n can be zero, λ is always positive (see the assumptions below). That is, the pressure equation is elliptic, and the saturation and concentration equations are parabolic. This model has been analyzed from the computational point of view using finite elements in [9, 10, 11].

3.3. The major result. The assumptions on the physical data are stated below; some of the assumptions in the previous section are repeated for completedness.

(B1) Assume that $\Omega \subset \Re^d$ is a multiply connected domain with Lipschitz boundary Γ, $\Gamma = \Gamma_1^p \cup \Gamma_2^p = \Gamma_1^\theta \cup \Gamma_2^\theta = \Gamma_1^c \cup \Gamma_2^c$, $\Gamma_1^p \cap \Gamma_2^p = \Gamma_1^\theta \cap \Gamma_2^\theta = \Gamma_1^c \cap \Gamma_2^c = \emptyset$, each Γ_i^p, Γ_i^θ, and Γ_i^c is a $(d-1)$-dimensional domain, and $\Gamma_2^p \subset \Gamma_2^\theta \cap \Gamma_2^c$.

(B2) Assume that $\phi \in L^\infty(\Omega)$, $\phi(x) \ge \phi_* > 0$, and $k(x)$ is a bounded, symmetric, and uniformly positive definite matrix, i.e.,

$$0 < k_* \le |\xi|^{-2} \sum_{i,j=1}^d k_{ij}(x)\xi_i\xi_j \le k^* < \infty, \quad x\in\Omega,\ \xi\ne 0\in\Re^d.$$

(B3) The diffusion-dispersion term in (3.5) is assumed to be Fickian in form with the coefficient given by [7]

$$D_\alpha(u_\alpha) = d_{m_\alpha} I + |u_\alpha|\big(d_{l_\alpha}E(u_\alpha) + d_{t_\alpha}E^\perp(u_\alpha)\big), \quad \alpha = w,n,$$

where $d_{m_\alpha} > 0$ is the molecular diffusion coefficient, d_{l_α} and d_{t_α} are the longitudinal and transverse dispersion coefficients, respectively, for the α phase, the matrix $E(u_\alpha)$ is the projection along the direction of flow determined by

$$E(u_\alpha) = \left(\frac{u_{\alpha,i}u_{\alpha,j}}{|u_\alpha|^2}\right), \quad |u_\alpha| = \sqrt{u_{\alpha,1}^2 + \cdots + u_{\alpha,d}^2}, \quad u_\alpha = (u_{\alpha,1},\ldots,u_{\alpha,d}),$$

and $E^\perp(u_\alpha) = I - E(u_\alpha)$.

(B4) $\lambda_\alpha(x,s,c)$ is measurable in $x\in\Omega$ and continuous in s, $c\in[0,1]$ and satisfies that $\lambda_w(0,c)=0$, $\lambda_w(s,c)>0$ for $s>0$, $\lambda_n(1,c)=0$, $\lambda_n(s,0)>0$ for $s<1$, and $0<\lambda_*\le\lambda(x,s,c)\le\lambda^*<\infty$, $x\in\Omega$, s, $c\in[0,1]$.

(B5) Assume that $0<\theta^*\in H^1(\Omega)$ and that $\mathcal{S}:\{(x,\theta):x\in\Omega,\,0\le\theta\le\theta^*(x)\}\to[0,1]$ is measurable in x, continuous and strictly monotone increasing in θ, and satisfies that $\mathcal{S}(x,0)=0$ and $\mathcal{S}(x,\theta^*(x))=1$.

(B6) Suppose that γ_1, γ_2, γ_3, q, q_w, R, and $\hat q$ are continuous in s and c and the following norms are bounded:

$$|||\gamma_1|||_{L^\infty(J;L^2(\Omega))},\quad |||\gamma_2|||_{L^2(\Omega_T)},\quad |||\varphi_1|||_{L^\infty(J;H^{-1/2}(\Gamma_1^p))},\quad |||R|||_{L^2(\Omega_T)},$$
$$|||q_w|||_{L^2(J;H^{-1}(\Omega))},\quad |||\gamma_3|||_{L^2(\Omega_T)},\quad |||\varphi_3|||_{L^2(J;H^{-1/2}(\Gamma_1^\theta))},$$
$$|||\varphi_5|||_{L^2(J;H^{-1/2}(\Gamma_1^c))},\quad |||\hat q|||_{L^2(J;H^{-1}(\Omega))},\quad |||q|||_{L^\infty(J;L^2(\Omega))},$$

where for $v=v(x,s,c)$,

$$|||v||| = \Big\| \sup_{s,c\in[0,1]} |v(x,s,c)| \Big\|$$

for any given norm. Also, for $s,c\in[0,1]$, assume that

$$\gamma_2(0,c)=0,\quad \gamma_1(1,c)-\gamma_2(1,c)=\lambda(1,c)\nabla\theta^*\quad \text{on } \Omega_T,$$
$$\varphi_1(1,c)\le 0,\quad \varphi_3(0,c)\le 0,\quad \varphi_3(1,c)\ge 0\quad \text{on } \Gamma_1^\theta,$$
$$\varphi_1(s,0)\le 0,\quad \varphi_1(s,1)\le 0,\quad \varphi_5(s,0)\le 0,\quad \varphi_5(s,1)\ge 0\quad \text{on } \Gamma_1^c.$$

(B7) Assume that $\partial_t\varphi_4$, $\partial_t\varphi_6\in L^1(\Omega_T)$ and

$$\varphi_2\in L^\infty(J;H^1(\Omega)),\ \varphi_4\in L^2(J;H^1(\Omega)),\ \varphi_6\in L^2(J;W^{1,4}(\Omega)),$$
$$0\le\varphi_4(x,t)\le\theta^*(x),\ 0\le\varphi_6(x,t)\le 1\quad \text{a.e. on } \Omega_T.$$

(B8) In the case of $\Gamma_2^p=\emptyset$ and $a_1\equiv 0$, q and φ_1 are independent of s and c, and

$$\int_{\Gamma_1^p}\varphi_1 \mathrm{d}\sigma = \int_\Omega q\mathrm{d}x.$$

(B9) There is a subset $\Gamma_{1,*}^p\subset\Gamma_1^p$ (with nonzero measure only if $\Gamma_2^p=\emptyset$ and $a_1\not\equiv 0$) such that $a_1\ge a_{1,*}>0$ on $\Gamma_{1,*}^p\times J$.

(B10) $a_i\ge 0$, a_i is continuous in s and c, the norm $|||a_i|||_{L^\infty(\Omega_T)}$ is bounded, $i=1,2,3$, and

$$a_1(1,c)=0\quad \text{on } \Gamma_1^\theta,\quad a_1(s,0)=a_1(s,1)=0\quad \text{on } \Gamma_1^c.$$

(B11) Assume that $q_w(0,c)\ge 0$, $q_n(1,c)=q(1,c)-q_w(1,c)\ge 0$, $q(s,0)\ge 0$, $q(s,1)\ge 0$, $\hat q(s,0)\ge 0$, $R(s)\ge 0$, and $R(s)-\hat c\hat q(s,1)\ge 0$ in Ω_T for $s,c\in[0,1]$, and $0\le\hat c\le 1$ a.e. on Ω_T.

(B12) Let θ_0, $c_0\in L^2(\Omega)$ satisfy $0\le\theta_0\le\theta^*(x)$ and $0\le c_0\le 1$ a.e. on Ω.

Define the spaces

$$V=\left\{v\in H^1(\Omega): v|_{\Gamma_2^p}=0;\ \text{if } \Gamma_2^p=\emptyset \text{ and } a_1\equiv 0,\ \text{then } \int_\Omega v\mathrm{d}x=0\right\},$$
$$W=\{v\in H^1(\Omega): v|_{\Gamma_2^\theta}=0\},$$
$$\Lambda=\{v\in H^1(\Omega): v|_{\Gamma_2^c}=0\}.$$

DEFINITION 3.1. *A weak solution of the system in* (3.14)–(3.18) *is a triple of functions* (p, θ, c) *with* $p \in L^\infty(J; V)+\varphi_2$, $\theta \in L^2(J; W)+\varphi_4$, $c \in L^2(J; \Lambda(s, u_w, u_n))+\varphi_6$ *such that*

$$\phi\partial_t s \in L^2(J; W^*), \quad \phi\partial_t c \in L^2(J; \Lambda^*(s, u_w, u_n)),$$

$$0 \le \theta(x,t) \le \theta^*(x), \quad 0 \le c(x,t) \le 1 \quad \text{a.e. on } \Omega_T,$$

$$s = \mathcal{S}(\theta),$$

$$\begin{aligned}(k\{\lambda(s,c)\nabla p + \gamma_1(s,c)\}, \nabla v) + (a_1(s,c)p, v)_{\Gamma_1^p} \\ = (q(s,c), v) - (\varphi_1(s,c), v)_{\Gamma_1^p} \quad \forall v \in L^\infty(J; V),\end{aligned}$$

$$\begin{aligned}&\int_J \langle \phi\partial_t s, v\rangle\, dt + \int_J (k\{\lambda(s,c)\nabla\theta + \lambda_w(s,c)\nabla p + \gamma_2(s,c)\}, \nabla v)\, dt \\ &+ \int_J (a_2(s,c)\theta, v)_{\Gamma_1^\theta}\, dt = \int_J (q_w(s,c), v)\, dt - \int_J (\varphi_3(s,c), v)_{\Gamma_1^\theta}\, dt \quad \forall v \in L^2(J; W),\end{aligned}$$

$$\begin{aligned}\int_J \langle \phi\partial_t s, v\rangle\, dt + \int_J (\phi(s - s_0), \partial_t v)\, dt = 0 \\ \forall v \in L^2(J; W) \cap W^{1,1}(J; L^1(\Omega)), \quad v(x,T) = 0,\end{aligned}$$

$$\begin{aligned}&\int_J \langle \phi\partial_t c, v\rangle\, dt + \int_J (D(s, u_w, u_n)\nabla c - uc, \nabla v)\, dt + \int_J (R(s)c, v)\, dt \\ &+ \int_J (a_3(s,c)c, v)_{\Gamma_1^c}\, dt = \int_J (\hat{c}\hat{q}(s,c), v)\, dt - \int_J (\varphi_5(s,c), v)_{\Gamma_1^c}\, dt \\ &\qquad \forall v \in L^2(J; \Lambda(s, u_w, u_n)),\end{aligned}$$

$$\begin{aligned}\int_J \langle \phi\partial_t c, v\rangle\, dt + \int_J (\phi(c - c_0), \partial_t v)\, dt = 0 \\ \forall v \in L^2(J; \Lambda(s, u_w, u_n)) \cap W^{1,1}(J; L^1(\Omega)), \quad v(x,T) = 0,\end{aligned}$$

where $s_0 = \mathcal{S}(\theta_0)$*;* u*,* u_w*, and* u_n *are given in* (3.15)*; and*

$$\Lambda(s, u_w, u_n) = \{v \in \Lambda : (D(s, u_w, u_n)\nabla v, \nabla v) < \infty\}.$$

THEOREM 3.2. *Under assumptions* (B1)–(B12)*, the system in* (3.14)–(3.18) *has a weak solution in the sense of Definition* 3.1.

3.4. Proof of the major result. With the same notation as in section 2.3, the discrete counterpart of Definition 3.1 is as follows: Find $p^h \in l_h(V) + \varphi_2^h$, $\theta^h \in l_h(W) + \varphi_4^h$, and $c^h \in l_h(\Lambda(s^h, u_w^h, u_n^h)) + \varphi_6^h$ such that

$$\begin{aligned}(k\{\lambda(s^h, c^h)\nabla p^h + \gamma_1(s^h, c^h)\}, \nabla v) + (a_1(s^h, c^h)p^h, v)_{\Gamma_1^p} \\ = (q(s^h, c^h), v) - (\varphi_1(s^h, c^h), v)_{\Gamma_1^p} \quad \forall v \in l_h(V),\end{aligned} \tag{3.19}$$

$$
\begin{aligned}
&\int_J \left(\phi\partial^{-h}s^h, v\right)\mathrm{d}t + \int_J (k\{\lambda(s^h,c^h)\nabla\theta^h + \lambda_w(s^h,c^h)\nabla p^h\}, \nabla v)\mathrm{d}t \\
&+ \int_J (k\gamma_2(s^h,c^h), \nabla v)\mathrm{d}t + \int_J (a_2(s^h,c^h)\theta^h, v)_{\Gamma_1^\theta}\mathrm{d}t \\
&\quad = \int_J (q_w(s^h,c^h), v)\mathrm{d}t - \int_J (\varphi_3(s^h,c^h), v)_{\Gamma_1^\theta}\mathrm{d}t \quad \forall v \in l_h(W),
\end{aligned} \tag{3.20}
$$

and

$$
\begin{aligned}
&\int_J \left(\phi\partial^{-h}c^h, v\right)\mathrm{d}t + \int_J (D(s^h,u_w^h,u_n^h)\nabla c^h - u^h c^h, \nabla v)\mathrm{d}t \\
&+ \int_J (R(s^h)c^h, v)\mathrm{d}t + \int_J (a_3(s^h,c^h)c^h, v)_{\Gamma_1^c}\mathrm{d}t \\
&\quad = \int_J (\hat{c}^h\hat{q}(s^h,c^h), v)\mathrm{d}t - \int_J (\varphi_5(s^h,c^h), v)_{\Gamma_1^c}\mathrm{d}t \quad \forall v \in l_h(\Lambda(s^h,u_w^h,u_n^h)),
\end{aligned} \tag{3.21}
$$

with u^h, u_w^h, and u_n^h being defined as in (3.15) and $s^h = s_0$ and $c^h = c_0$ for $t < 0$.

LEMMA 3.3. *With $\alpha = w, n$, it holds that*

$$
\begin{aligned}
d_{m_\alpha} + \min(d_{l_\alpha}, d_{t_\alpha})|u_\alpha| \le& \phi^{-1}|\xi|^{-2}\sum_{i,j=1}^d D_{\alpha,ij}(u_\alpha)\xi_i\xi_j \\
&\le d_{m_\alpha} + \max(d_{l_\alpha}, d_{t_\alpha})|u_\alpha|, \quad \xi \neq 0 \in \Re^d,
\end{aligned}
$$

and for $s \in [0,1]$

$$
\begin{aligned}
&\min(d_{m_w}, d_{m_n}) + \min(d_{l_w}, d_{t_w}, d_{l_n}, d_{t_n})(s|u_w| + (1-s)|u_n|) \\
&\le \phi^{-1}|\xi|^{-2}\sum_{i,j=1}^d D_{ij}(s,u_w,u_n)\xi_i\xi_j \\
&\le \max(d_{m_w}, d_{m_n}) + \max(d_{l_w}, d_{t_w}, d_{l_n}, d_{t_n})(s|u_w| + (1-s)|u_n|), \quad \xi \neq 0 \in \Re^d.
\end{aligned}
$$

The first part follows from the definition of $D_\alpha(u_\alpha)$, while the second part comes from the definition of $D(s,u_w,u_n)$.

LEMMA 3.4. *For $h > 0$ small enough, the discrete scheme has a solution such that*

$$
0 \le \theta^h(x,t) \le \theta^*(x), \quad 0 \le c^h(x,t) \le 1 \quad \textit{a.e. on} \quad \Omega_T. \tag{3.22}
$$

This lemma will be shown in the next subsection.

LEMMA 3.5. *The solution to the discrete scheme also satisfies*

$$
\begin{aligned}
&\|p^h\|_{L^\infty(J;H^1(\Omega))} + \|\theta^h\|_{L^2(J;H^1(\Omega))} + \|c^h\|_{L^2(J;H^1(\Omega))} \\
&\quad + \|D^{1/2}(s^h,u_w^h,u_n^h)\nabla c^h\|_{L^2(\Omega_T)} \le C.
\end{aligned} \tag{3.23}
$$

Proof. The bound on p^h can be obtained as in Lemma 2.5 using (3.19); the estimate on θ^h also can be seen from (3.20) using an argument similar to that for proving c^h in Lemma 2.5 (also see [6]). It suffices to obtain a bound on c^h from (3.21). Again, attention is paid to the transport and diffusion-dispersion terms; other terms can be estimated similarly to the method in the proof of Lemma 2.5.

Take $v = c^h - \varphi_6^h \in l_h(\Lambda(s^h, u_w^h, u_n^h))$ in (3.21) to see that

$$\int_J \left(\phi \partial^{-h} c^h, c^h - \varphi_6^h\right) \mathrm{d}t + \int_J (D(s^h, u_w^h, u_n^h)\nabla c^h - u^h c^h, \nabla[c^h - \varphi_6^h])\mathrm{d}t$$
$$+ \int_J (R(s^h)c^h, c^h - \varphi_6^h)\mathrm{d}t + \int_J (a_3(s^h, c^h)c^h, c^h - \varphi_6^h)_{\Gamma_1^c}\mathrm{d}t$$
$$= \int_J (\hat{c}^h \hat{q}(s^h, c^h), c^h - \varphi_6^h)\mathrm{d}t - \int_J (\varphi_5(s^h, c^h), c^h - \varphi_6^h)_{\Gamma_1^c}\mathrm{d}t.$$

The transport and diffusion-dispersion terms can be estimated as follows. By Lemma 3.3, the definition of u_w^h and u_n^h, and the above bound on p^h and θ^h, we have

$$\begin{aligned}
&|(D(s^h, u_w^h, u_n^h)\nabla c^h, \nabla \varphi_6^h)| \\
&\quad \le \epsilon(D(s^h, u_w^h, u_n^h)\nabla c^h, \nabla c^h) + C(D(s^h, u_w^h, u_n^h)\nabla \varphi_6^h, \nabla \varphi_6^h) \\
&\quad \le \epsilon(D(s^h, u_w^h, u_n^h)\nabla c^h, \nabla c^h) + C\|\nabla \varphi_6^h\|_{L^4(\Omega)}^2.
\end{aligned}$$

Also, it follows from Lemmas 3.3 and 3.4 that

$$|(u^h c^h, \nabla c^h)| \le \|u^h\|_{L^2(\Omega)}\|\nabla c^h\|_{L^2(\Omega)} \le \epsilon(D(s^h, u_w^h, u_n^h)\nabla c^h, \nabla c^h) + C\|u^h\|_{L^2(\Omega)}^2$$

and

$$|(u^h c^h, \nabla \varphi_6^h)| \le C\left(\|u^h\|_{L^2(\Omega)}^2 + \|\nabla \varphi_6^h\|_{L^2(\Omega)}^2\right).$$

Apply these estimates and the argument used in the proof of Lemma 2.5 to yield the desired result. □

From this lemma, we have the following counterpart of Corollary 2.6; the proof is also the same.

COROLLARY 3.6. *For any $2 \le r < \infty$, for a subsequence $p^h \rightharpoonup p$ weakly in $L^r(J; H^1(\Omega))$ and $\theta^h \rightharpoonup \theta$ and $c^h \rightharpoonup c$ weakly in $L^2(J; H^1(\Omega))$. Furthermore, $p \in L^\infty(J; V) + \varphi_2$, $\theta \in L^2(J; W) + \varphi_4$, $c \in L^2(J; \Lambda) + \varphi_6$, and*

$$0 \le \theta(x,t) \le \theta^*(x), \quad 0 \le c(x,t) \le 1 \quad \text{a.e. on} \quad \Omega_T.$$

LEMMA 3.7. *There is a subsequence such that $\theta^h \to \theta$ and $c^h \to c$ strongly in $L^2(\Omega_T)$.*

This lemma can be shown in the same manner as in Lemma 2.7 (also see [3, 6]). From it, we have the next corollary, whose proof is the same as in Corollary 2.8. The pointwise convergence for $\{s^h\}$ follows from the continuity of $\mathcal{S}(\theta)$ in θ.

COROLLARY 3.8. *There is a subsequence such that $\theta^h \to \theta$ and $c^h \to c$ strongly in $L^2(J; H^{1-\pi}(\Omega))$ and $L^2(J; H^{1/2-\pi}(\partial\Omega))$ for any $0 < \pi < 1/2$, and $s^h \to s$ and $c^h \to c$ pointwise a.e. on Ω_T.*

Applying Corollaries 3.6 and 3.8 and the same techniques as in the proof of Theorem 2.2, Theorem 3.2 can be shown as before. Thus, it remains to prove Lemma 3.4, which will be carried out in the next subsection.

3.5. Proof of Lemma 3.4. Again, Lemma 3.4 is purely an elliptic result and will follow from the next proposition. For notational convenience the superscript h is omitted below. All functions of s and c are extended constantly outside $[0, 1]$ except $\mathcal{S}$, which is extended as follows [6]:

$$s = \text{extended } \mathcal{S}(\theta) = \begin{cases} \theta & \text{for} \quad \theta < 0, \\ \mathcal{S}(x, \theta) & \text{for} \quad 0 \le \theta \le \theta^*(x), \\ \theta + 1 - \theta^*(x) & \text{for} \quad \theta^*(x) < \theta. \end{cases}$$

PROPOSITION 3.9. *In addition to assumptions* (B1)–(B12)*, suppose that* $0 < \eta_* \le \eta_1(x) \in L^\infty(\Omega)$*,* $0 \le \eta_2(x) \le \eta_1(x)$*,* $0 < \zeta_* \le \zeta_1(x) \in L^\infty(\Omega)$*, and* $0 \le \zeta_2(x) \le \zeta_1(x)$*. Then, for* η_* *and* ζ_* *sufficiently big, the following problem has a weak solution* $(p, \theta, c) \in (V + \varphi_2) \times (W + \varphi_4) \times (\Lambda(s, u_w, u_n) + \varphi_6)$*:*

$$\begin{aligned}(k\{\lambda(s,c)\nabla p + \gamma_1(s,c)\}, \nabla v) + (a_1(s,c)p, v)_{\Gamma_1^p}\\ = (q(s,c), v) - (\varphi_1(s,c), v)_{\Gamma_1^p} \quad \forall v \in V,\end{aligned} \tag{3.24}$$

$$\begin{aligned}&(\eta_1 s, v) + (k\{\lambda(s,c)\nabla\theta + \lambda_w(s,c)\nabla p\}, \nabla v)\\ &\quad + (k\gamma_2(s,c), \nabla v) + (a_2(s,c)\theta, v)_{\Gamma_1^\theta}\\ &\quad = (q_w(s,c), v) - (\varphi_3(s,c), v)_{\Gamma_1^\theta} + (\eta_2, v) \quad \forall v \in W,\end{aligned} \tag{3.25}$$

$$\begin{aligned}&(\zeta_1 c, v) + (D(s, u_w, u_n)\nabla c - uc, \nabla v)\\ &\quad + (R(s)c, v) + (a_3(s,c)c, v)_{\Gamma_1^c}\\ &\quad = (\hat{c}\hat{q}(s,c), v) - (\varphi_5(s,c), v)_{\Gamma_1^c} + (\zeta_2, v) \quad \forall v \in \Lambda(s, u_w, u_n),\end{aligned} \tag{3.26}$$

and

$$0 \le \theta(x,t) \le \theta^*(x), \quad 0 \le c(x,t) \le 1 \quad \text{a.e. on } \Omega_T, \tag{3.27}$$

where u*,* u_w*, and* u_n *are given as in* (3.15)*.*

Proof. Let $\{v_i^j\}_{i=1}^\infty$ $(j = 1, 2, 3)$ be bases for V, W, and Λ, respectively, and set $V_m = \text{span}\{v_1^1, \dots, v_m^1\}$, $W_m = \text{span}\{v_1^2, \dots, v_m^2\}$, and $\Lambda_m = \text{span}\{v_1^3, \dots, v_m^3\}$. With V_m, W_m, and Λ_m replacing V, W, and Λ in (3.24)–(3.26), respectively, we again obtain a Galerkin procedure.

For $v^j = \sum_{i=1}^m \beta_i^j v_i^j$, $j = 1, 2, 3$, we introduce the mapping $\Phi_m : \Re^{3m} \to \Re^{3m}$ by

$$\Phi_m \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^3 \end{pmatrix} = \begin{pmatrix} \hat{\beta}^1 \\ \hat{\beta}^2 \\ \hat{\beta}^3 \end{pmatrix},$$

where

$$\begin{aligned}\hat{\beta}_i^1 =& (k\{\lambda(\sigma, v^3 + \varphi_6)\nabla(v^1 + \varphi_2) + \gamma_1(\sigma, v^3 + \varphi_6)\}, \nabla v_i^1)\\ &+ (a_1(\sigma, v^3 + \varphi_6)(v^1 + \varphi_2), v_i^1)_{\Gamma_1^p} - (q(\sigma, v^3 + \varphi_6), v_i^1) + (\varphi_1(\sigma, v^3 + \varphi_6), v_i^1)_{\Gamma_1^p},\\ \hat{\beta}_i^2 =& \big(\eta_1\sigma, v_i^2\big) + (k\{\lambda(\sigma, v^3 + \varphi_6)\nabla(v^2 + \varphi_4) + \lambda_w(\sigma, v^3 + \varphi_6)\nabla(v^1 + \varphi_2)\}, \nabla v_i^2)\\ &+ (k\gamma_2(\sigma, v^3 + \varphi_6), \nabla v_i^2) + (a_2(\sigma, v^3 + \varphi_6)\theta, v_i^2)_{\Gamma_1^\theta}\\ &- (q_w(\sigma, v^3 + \varphi_6), v_i^2) + (\varphi_3(\sigma, v^3 + \varphi_6), v_i^2)_{\Gamma_1^\theta} - (\eta_2, v_i^2),\\ \hat{\beta}_i^3 =& \big(\zeta_1(v^3 + \varphi_6), v_i^3\big) + (D(\sigma, \hat{u}_w, \hat{u}_n)\nabla(v^3 + \varphi_6) - \hat{u}(v^3 + \varphi_6), \nabla v_i^3)\\ &+ (R(\sigma)(v^3 + \varphi_6), v_i^3) + (a_3(\sigma, v^3 + \varphi_6)(v^3 + \varphi_6), v_i^3)_{\Gamma_1^c}\\ &- (\hat{c}\hat{q}(\sigma, v^3 + \varphi_6), v_i^3) + (\varphi_5(\sigma, v^3 + \varphi_6), v_i^3)_{\Gamma_1^c} - (\zeta_2, v_i^3),\end{aligned}$$

where

$$\begin{aligned}&\sigma = \mathcal{S}(v^2 + \varphi_4),\\ &u = -k(\lambda(\sigma, v^3 + \varphi_6)\nabla(v^1 + \varphi_2) + \gamma_1(\sigma, v^3 + \varphi_6)),\\ &u_w = -k\big(\lambda(\sigma, v^3 + \varphi_6)\nabla(v^2 + \varphi_4) + \lambda_w(\sigma, v^3 + \varphi_6)\nabla(v^1 + \varphi_2) + \gamma_2(\sigma, v^3 + \varphi_6)\big),\\ &u_n = k\big(\lambda(\sigma, v^3 + \varphi_6)\nabla(v^2 + \varphi_4) - \lambda_n(\sigma, v^3 + \varphi_6)\nabla(v^1 + \varphi_2) + \gamma_3(\sigma, (v^1 + \varphi_2))\big),\\ &\hat{\xi} = m\xi/(m + |\xi|), \quad \xi = u, u_w, u_n.\end{aligned}$$

Again, we remark that to handle the difficulty associated with the transport and diffusion-dispersion terms, we have introduced $\hat{u}$, $\hat{u}_w$, and $\hat{u}_n$ above. By the assumptions (B1)–(B12), Φ_m is continuous, and in the same fashion as in the proof of Proposition 2.9, we obtain

$$\Phi_m \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^3 \end{pmatrix} \cdot \begin{pmatrix} \beta^1 \\ \beta^2 \\ \beta^3 \end{pmatrix} \geq C_1(m)\{\|v^1\|^2_{H^1(\Omega)} + \|v^2\|^2_{H^1(\Omega)} + \|v^3\|^2_{H^1(\Omega)}\} - C,$$

which is strictly positive for $|\beta^1| + |\beta^2|$ sufficiently big. Consequently, Φ_m has a zero; i.e., there is a solution to the Galerkin approximation with $\hat{u}$ (respectively, $\hat{u}_w$ and $\hat{u}_n$) replacing u (respectively, u_w and u_n) for each m.

The rest of the proof is completed with a standard maximum principle argument on (3.25) and (3.26). First, take $v = \theta^- = \min(\theta, 0) \in W$ in (3.25) to see that

$$\begin{aligned}(\eta_1 s - \eta_2, \theta^-) = &-(k\lambda(s,c)\nabla\theta, \nabla\theta^-) - (k\lambda_w(s,c)\nabla p, \nabla\theta^-) - (k\gamma_2(s,c), \nabla\theta^-) \\ &-(a_2(s,c)\theta, \theta^-)_{\Gamma_1^\theta} + (q_w(s,c), \theta^-) - (\varphi_3(s,c), \theta^-)_{\Gamma_1^\theta} \leq 0\end{aligned}$$

by assumptions (B4), (B6), and (B11). This implies that $\theta^- = 0$ a.e. on Ω_T by the definition of $\mathcal{S}$ provided η_* is sufficiently large; i.e., $\theta \geq 0$ a.e. on Ω_T. Now, with $v = (\theta - \theta^*)^+ = \max(\theta - \theta^*, 0) \in W$ in (3.25) and use (3.24) ($(\theta - \theta^*)^+ \in V$ by (B1); if $\Gamma_2^p = \emptyset$ and $a_1 \equiv 0$, consider $v - \int_\Omega v\mathrm{d}x$) we see that

$$\begin{aligned}(\eta_1 s - \eta_2, (\theta - \theta^*)^+) = &-(k\{\lambda(s,c)\nabla\theta - \lambda_n(s,c)\nabla p - \gamma_1(s,c) + \gamma_2(s,c)\}, \nabla(\theta - \theta^*)^+) \\ &+(a_1(s,c)p + \varphi_1(s,c) - a_2(s,c)\theta - \varphi_3(s,c), (\theta - \theta^*)^+)_{\Gamma_1^\theta} \\ &-(q_n(s,c), (\theta - \theta^*)^+) \leq 0\end{aligned}$$

by assumptions (B4), (B6), (B10), and (B11), from which we conclude that $\theta \leq \theta^*$. The second part in (3.27) can be shown as in the proof of Proposition 2.9. Thus, the proof is complete. □

REFERENCES

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

[2] H. W. Alt and E. di Benedetto, *Nonsteady flow of water and oil through inhomogeneous porous media*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 12 (1985), pp. 335–392.

[3] H. W. Alt and S. Luckhaus, *Quasilinear elliptic-parabolic differential equations*, Math. Z., 183 (1983), pp. 311–341.

[4] S. N. Antoncev, *On the solvability of boundary value problems for degenerate two-phase porous flow equations*, Dinamika Sploshn. Sredy Vyp., 10 (1972), pp. 28–53 (in Russian).

[5] S. N. Antoncev, A. V. Kazhikhov, and V. N. Monakhov, *Boundary-Value Problems in the Mechanics of Nonuniform Fluids*, Stud. Math. Appl. 22, North–Holland, Amsterdam, The Netherlands, 1990.

[6] T. J. Arbogast, *The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow*, Nonlinear Anal., 19 (1992), pp. 1009–1031.

[7] J. Bear, *Dynamics of Fluids in Porous Media*, Dover, New York, 1972.

[8] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation*, North–Holland, Amsterdam, The Netherlands, 1978.

[9] Z. Chen, M. Espedal, and R. Ewing, *Continuous-time finite element analysis of multiphase flow in groundwater hydrology*, Appl. Math., 40 (1995), pp. 203–226.

[10] Z. Chen and R. Ewing, *Fully discrete finite element analysis of multiphase flow in groundwater hydrology*, SIAM J. Numer. Anal., 34 (1997), pp. 2228–2253.

[11] Z. Chen, R. Ewing, and M. Espedal, *Multiphase flow simulation with various boundary conditions*, in Numerical Methods in Water Resources, 2, A. Peters, et al., eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 1994, pp. 925–932.

[12] J. Douglas, Jr., R. Ewing, and M. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Anal. Numér., 17 (1983), pp. 17–33.

[13] J. Douglas, Jr., and J. Robert, *Numerical methods for a model for compressible miscible displacement in porous media*, Math. Comp., 41 (1983), pp. 441–459.

[14] X. Feng, *On existence and uniqueness for a coupled system modeling miscible displacement in porous media*, J. Math. Anal. Appl., 194 (1995), pp. 441–469.

[15] D. Kroener and S. Luckhaus, *Flow of oil and water in a porous medium*, J. Differential Equations, 55 (1984), pp. 276–288.

[16] S. N. Kružkov and S. M. Sukorjanskiĭ, *Boundary problems for systems of equations of two-phase porous flow type; statement of the problems, questions of solvability, justification of approximate methods*, Math. USSR Sbornik, 33 (1977), pp. 62–80.

[17] O. A. Ladyzenskaja, V. A. Solonniko, and N. N. Uralćeva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, American Mathematical Society, Providence, RI, 1968.

[18] A. Mikelić, *Mathematical theory of stationary miscible filtration*, J. Differential Equations, 90 (1991), pp. 186–202.

[19] D. W. Peaceman, *Fundamentals of Numerical Reservoir Simulation*, Elsevier, New York, 1977.

[20] P. H. Sammon, *Numerical approximations for a miscible displacement in porous media*, SIAM J. Numer. Anal., 23 (1986), pp. 505–542.

SIAM J. MATH. ANAL.
Vol. 30, No. 2, pp. 454–467

SINGULARITIES OF DISTRIBUTIONS VIA THE WAVELET TRANSFORM*

JAIME NAVARRO†

Abstract. For a given distribution u in $\mathcal{S}'(\mathbf{R}^2)$, a wavelet transform of u with respect to an admissible function is defined in such a way that the wavelet transform of u yields a function on phase space whose high-frequency singularities are precisely the elements in the wave front set of u.

Key words. admissible functions, wavelet transform, wave front set

AMS subject classifications. 42A38, 44A05

PII. S0036141095292494

Introduction. The theory of distributions, or generalized functions, is a central tool in analysis, mathematical physics, and applied mathematics. A key element in distribution theory is the notion of the wave front set of a distribution. This set is a subset of phase space whose elements are pairs of locations and associated directions at which the distribution fails to be smooth. The wave front set plays several major roles in the theory of distributions, two of which are the following. First, the "pointwise" product of two distributions can be defined only when their wave front sets have empty intersection; knowledge of wave front sets proves crucial in discussions of singular solutions to certain nonlinear differential equations. Second, the wave front set of a distribution solution to a linear partial differential equation is invariant under flow of phase space that is associated with the partial differential operator; knowledge of the wave front set for Cauchy data allows very precise determination of the propagation of singularities in the corresponding solution.

Despite the utility and power of the concept, the definition of the wave front set is somewhat indirect. Suppose, for illustration purposes, that we wish to find the C^∞ wave front set for a distribution u that is defined pointwise (i.e., is an ordinary function) on $\mathbf{R}^n$.

Recall that a C^∞ function of compact support is characterized by the fact that its Fourier transform vanishes faster at large values of its argument k than the reciprocal of any polynomial in k. To determine whether the pair (x, k), where x is a point in $\mathbf{R}^n$ and k is a direction in the cotangent space to $\mathbf{R}^n$ at x, is in the C^∞ wave front set of u, we first localize u by multiplying it by a C^∞ function ϕ with compact support containing x. We then inspect the Fourier transform of ϕu in the direction k. If there is a bounded sequence of such cutoff functions ϕ whose supports converge to x for each of which the Fourier transform of ϕu in the direction k fails to fall off faster than the reciprocal of every polynomial in $|k|$, then the pair (x, k) is said to be in the wave front set for u. In this way, the wave front set specifies the directions k along which u fails to be smooth, at various points x.

*Received by the editors September 27, 1995; accepted for publication (in revised form) July 21, 1998; published electronically February 2, 1999.
http://www.siam.org/journals/sima/30-2/29249.html
†Universidad Autónoma Metropolitana, Unidad Azcapotzalco, Division de Ciencias Básicas, Apdo Postal 16-306, México D.F. 02000, México (jnfu@hp9000a1.uam.mx).

The projection of the C^∞ wave front set onto configuration space $\boldsymbol{R^n}$ is the set of points x at which u fails to be C^∞. But the wave front set itself has not been identified as the set of singularities of any distribution on phase space. Especially in applications to partial differential equations, where many concepts are formulated using functions on phase space, the definition of the wave front subset of phase space without reference to an explicit function on phase space seems conceptually incomplete.

This paper supplies precisely this missing conceptual ingredient by formulating an explicit transform that, applied to any distribution in $\mathcal{S}'(\mathbf{R}^2)$, yields a function on phase space whose singularities are the wave front set of a distribution.

The definition of the wave front set of a distribution involves the high-frequency behavior of each member of a sequence of windowed Fourier transforms of the distribution. Finding singularities of a function through use of the windowed Fourier transform is known to be difficult; this difficulty is reflected in the wave front set definition. In contrast, the wavelet transform is well suited to the direction of singularities and offers an alternative to the windowed Fourier transform [1, 4]. Furthermore, the wavelet transform of a distribution can be regarded as a function on phase space. Rephrasing the definition of the wave front set in terms of the windowed Fourier transform thus leads to the following characterization: The C^∞ wave front set of a distribution is precisely the set of high-frequency singularities of the wavelet transform of the distribution with respect to a suitably chosen C^∞ basic wavelet.

The continuous wavelet transform thus provides a simple explicit constructive characterization for the wave front set. It also holds promise of far greater utility. The invariance of the wave front set of a solution to a partial differential equation under the bicharacteristic flow generated by the symbol of the differential operator is a very powerful tool. It is likely that the wavelet transform of a solution exhibits simple behavior under the bicharacteristic flow on phase space, with the high-frequency limit invariant under that flow. It is conceivable that the wavelet transform will provide a canonical decomposition of initial data for hyperbolic partial differential equations into component portions that propagate along bicharacteristic curves.

In this paper we construct a mother wavelet in $\mathcal{S}(\mathbf{R}^2)$ and an irreducible group action with the property that the associated wavelet transform of a distribution in $\mathcal{S}'(\mathbf{R}^2)$ is singular along the wave front set of the distribution.

Singularities in direction (1, 0) are treated only in frequency space, but rotation can be used to investigate the wave front set in other directions. A mother wavelet in $\mathcal{S}(\mathbf{R}^2)$ and an irreducible group action are constructed with the property that the associated wavelet transform of a distribution in $\mathcal{S}'(\mathbf{R}^2)$ is singular along the wave front set of the distribution. First, the definition of the wavelet transform of distributions in $\mathcal{S}(\mathbf{R}^2)$ is given, and then the main result is stated, which relates the notion of the wave front set and the wavelet transform of distributions in $\mathcal{S}'(\mathbf{R}^2)$.

Let us begin by defining an irreducible group action on $L^2(\mathbf{R}^2)$.

NOTATIONS.

1. *For a in $\mathbf{R} \setminus \{0\}$, let* $M(a) = \begin{pmatrix} a^2 & 0 \\ 0 & a \end{pmatrix}$.

2. *Let $Q = \{z \in \mathbf{C}: \mid z \mid = 1\}$. Identify Q with $[0,1)$ by associating with $r \in [0,1)$ the complex number $\zeta = e^{2\pi i r} \in Q$.*

DEFINITION 1. *For h in $L^2(\mathbf{R}^2)$, define the following operators:*

$$(J_a h)(x) = \frac{1}{\sqrt{|\det\ M(a)|}} h(M(a)^{-1}x), \quad x \in \mathbf{R}^2, \quad a \in \mathbf{R} \setminus \{0\},$$

$$(T_b h)(x) = h(x-b), \quad x, b \in \mathbf{R}^2,$$

$$(E_b h)(x) = e^{2\pi i x\cdot b} h(x), \quad x, b \in \mathbf{R}^2,$$
$$(T_\tau h)(x) = h(x_1, x_2 - \tau), \quad x = (x_1, x_2) \in \mathbf{R}^2, \quad \tau \in \mathbf{R},$$
$$(E_\tau h)(x) = e^{2\pi i x_2 \tau} h(x), \quad x = (x_1, x_2) \in \mathbf{R}^2, \quad \tau \in \mathbf{R}.$$

DEFINITION 2. *Let* $G = \{(a, b, \tau, \zeta) : a \in \mathbf{R} \setminus \{0\}, b \in \mathbf{R}^2, \tau \in \mathbf{R}, \text{and } \zeta \in Q\}$. *For* $(a_1, b_1, \tau_1, \zeta_1)$ *and* $(a_2, b_2, \tau_2, \zeta_2)$ *in* G, *define*

$$(a_1, b_1, \tau_1, \zeta_1) \cdot (a_2, b_2, \tau_2, \zeta_2) = \left(a_1 a_2, b_1 + M(a_1) b_2, \tau_1 + \frac{\tau_2}{a_1}, \zeta_1 \zeta_2 e^{-2\pi i b_{12} \frac{\tau_2}{a_1}} \right),$$

where $b_1 = (b_{11}, b_{12})$ *and* $b_2 = (b_{21}, b_{22})$ *are in* $\mathbf{R}^2$.

Remark 1. G is a nonunimodular, locally compact topological group, with identity $(1, 0, 0, 1)$ and

$$(a, b, \tau, \zeta)^{-1} = (a^{-1}, -M(a)^{-1} b, -a\tau, \zeta^{-1} e^{-2\pi i b_2 \tau}),$$

where $b = (b_1, b_2)$ *is in* $\mathbf{R}^2$. Furthermore, $d(a, b, \tau, \zeta) = \frac{1}{|\det\ M(a)|}\, da\, db\, d\tau\, dr$ and $d_1(a, b, \tau, \zeta) = \frac{1}{|a|}\, da\, db\, d\tau\, dr$ are the left and right Haar measures, respectively, with $\zeta = e^{2\pi i r}$.

DEFINITION 3. *For* (a, b, τ, ζ) *in* G, *define the* 4*-parameter family of operators* $U(a, b, \tau, \zeta) = \zeta E_\tau T_b J_a$. $U(a, b, \tau, \zeta)$ *acts on the Hilbert space* $L^2(\mathbf{R}^2)$ *by*

$$(U(a, b, \tau, \zeta) h)(x) \equiv (\zeta E_\tau T_b J_a h)(x) = \zeta e^{2\pi i x_2 \tau} \frac{1}{\sqrt{|\det\ M(a)|}} h(M(a)^{-1}(x - b)).$$

DEFINITION 4. *For* (a, b, τ, ζ) *in* G, h *in* $\mathcal{S}(\mathbf{R}^2)$, *and* u *in* $\mathcal{S}'(\mathbf{R}^2)$, *define the wavelet transform of* u *with respect to* h *as*

$$(L_h u)(a, b, \tau, \zeta) = u[\overline{\zeta E_\tau T_b J_a h}].$$

The action of G on $\mathcal{S}(\mathbf{R}^2)$ has been chosen to meet several criteria that will be explained in the proof of Theorem 1.

The most important property of the action of G is that, as $a \to 0$, the function $(U(a, b, \tau, \zeta)h)(x)$ concentrates near $x = b$ in coordinate space, while its Fourier transform $(\widehat{U(a, b, \tau, \zeta)h})(k) = (\zeta\, T_\tau E_{-b} J_{1/a} \hat{h})(k)$ concentrates in cones around $(1, 0)$ in Fourier transform space.

The action in G on the Fourier transform $\hat{h}$ is primarily through the operators T_τ and $J_{1/a}$ that transform the arguments of $\hat{h}$. To discuss this aspect of the action of G, we introduce the map $\mathcal{V} : \mathbf{R}^2 \to \mathbf{R}^2$ such that for each (a, τ) in $(\mathbf{R} \setminus \{0\}) \times \mathbf{R}$, $\mathcal{V}(a, \tau)(k_1, k_2) \equiv (\frac{k_1}{a^2}, \frac{k_2}{a} + \tau)$.

It is also convenient to define the set $\triangle_\xi$ of parameters (a, τ) for which the image of a rectangle Ω covers the point ξ.

DEFINITION 5. *Let* $\xi = (\xi_1, \xi_2)$ *be in* $\mathbf{R}^+ \times \mathbf{R}$. *Let* $\Omega = [c_1, c_2] \times [d_1, d_2]$, *where* $0 < c_1 < c_2$ *and* $d_1 < 0 < d_2$. *Define* $\triangle_\xi = \{(a, \tau) \in (\mathbf{R} \setminus \{0\}) \times \mathbf{R} : \mathcal{V}(a, \tau)\Omega \ni \xi\}$.

LEMMA 1. *Let* $\xi = (\xi_1, \xi_2)$ *be in* $\mathbf{R}^+ \times \mathbf{R}$. *Suppose that* $\Omega = [c_1, c_2] \times [d_1, d_2]$, *where* $0 < c_1 < c_2$ *and* $d_1 < 0 < d_2$. *Then*

$$\triangle_\xi = \left\{ (a, \tau) \in (\mathbf{R} \setminus \{0\}) \times \mathbf{R} : \sqrt{\frac{c_1}{\xi_1}} \le |a| \le \sqrt{\frac{c_2}{\xi_1}} \quad \text{and} \quad \xi_2 - \frac{d_2}{|a|} \le \tau \le \xi_2 - \frac{d_1}{|a|} \right\}.$$

See the appendix for the proof.

Remark 2. It follows by Lemma 1 that if $\Omega = [c_1, c_2] \times [d_1, d_2]$, where $0 < c_1 < c_2$ and $d_1 < 0 < d_2$, then for each (a, τ) in $\triangle_\xi$,

$$\mathcal{V}(a,\tau)\Omega = \left\{(\eta_1, \eta_2) \in \mathbf{R}^+ \times \mathbf{R} : \frac{c_1}{a^2} \le \eta_1 \le \frac{c_2}{a^2} \text{ and } \frac{d_1}{|a|} + \tau \le \eta_2 \le \frac{d_2}{|a|} + \tau\right\}.$$

LEMMA 2. *Let $\Omega = [c_1, c_2] \times [d_1, d_2]$, where $0 < c_1 < c_2$ and $d_1 < 0 < d_2$. Let Γ_1 be an open cone, symmetric with respect to $(1, 0)$ with opening half-angle $\beta, 0 < \beta < \frac{\pi}{2}$. Then there is a nonempty open cone $\Gamma \subset \Gamma_1$, symmetric with respect to $(1, 0)$ with opening half-angle $\alpha, 0 < \alpha < \frac{\pi}{2}$, such that for any ξ in Γ with $|\xi|$ sufficiently large, $\mathcal{V}(a, \tau)\Omega \subset \Gamma_1$ for all (a, τ) in $\triangle_\xi$.*

See the appendix for the proof.

DEFINITION 6. *Let $\Omega = [c_1, c_2] \times [d_1, d_2]$, where $0 < c_1 < c_2$ and $d_1 < 0 < d_2$, and let $X \subset \mathbf{R}^2$. For u in $\mathcal{S}'(\mathbf{R}^2)$ and h in $\mathcal{S}(\mathbf{R}^2)$, define*

$$(L_h u)|_{(X, \triangle_\xi)} = \text{ Sup } \{|(L_h u)(a, b, \tau, \zeta)| : (a, \tau) \in \triangle_\xi, b \in X, \text{ and } \zeta \in Q\},$$

where $\triangle_\xi = \{(a, \tau) \in (\mathbf{R} \setminus \{0\}) \times \mathbf{R} : \mathcal{V}(a, \tau)\Omega \ni \xi\}$, and $\mathcal{V}(a, \tau)(k_1, k_2) = (\frac{k_1}{a^2}, \frac{k_2}{a} + \tau)$.

The following two definitions of the wave front set for distributions are due to Hörmander [5].

DEFINITION 7. *Let u be in $\mathcal{E}'(V)$, where V is an open set in $\mathbf{R}^2$. Define the cone $\sum(u)$ as the set of all η in $\mathbf{R}^2 \setminus \{0\}$ having no open conic neighborhood Γ in which $(1 + |\xi|)^N |\hat{u}(\xi)|$ is bounded for all positive integers N and all ξ in Γ.*

DEFINITION 8. *Let u be in $\mathcal{D}'(V)$, where V is an open set in $\mathbf{R}^2$. Then the closed subset of $V \times (\mathbf{R}^2 \setminus \{0\})$ defined by $WF(u) = \{(x, \xi) \in V \times (\mathbf{R}^2 \setminus \{0\}) : \xi \in \sum_x(u)\}$ is called the wave front set of u, where $\sum_x(u) = \bigcap_\phi \sum(\phi u), \phi \in \mathcal{D}(V)$, and $\phi(x) \neq 0$.*

The main result of this paper is now given.

THEOREM 1. *Let u be in $\mathcal{S}'(\mathbf{R}^2)$. Let h be a function such that $\hat{h} \in \mathcal{D}(\mathbf{R}^2)$ and* supp $\hat{h} \subset \Omega$, *where $\Omega = [c_1, c_2] \times [d_1, d_2]$, $0 < c_1 < c_2$ and $d_1 < 0 < d_2$, and h is not identically zero. Let $\xi = (\xi_1, \xi_2)$ be in $\mathbf{R}^+ \times \mathbf{R}$. Let $\triangle_\xi$ be as defined. Let $(x_0, (1, 0))$ be in $\mathbf{R}^2 \times \mathbf{R}^2$.*

(1) *Suppose that $(x_0, (1, 0)) \notin WF(u)$. Then there is an open neighborhood X of x_0 and an open conic neighborhood Γ of $(1, 0)$ such that $(L_h u)|_{(X, \triangle_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ.*

(2) *Conversely, suppose that there is a function $\psi \in \mathcal{D}(\mathbf{R}^2)$ with $\psi = 1$ in a neighborhood of x_0, an open neighborhood X of* supp ψ, *and an open conic neighborhood Γ of $(1, 0)$ such that $(L_h \psi u)|_{(X, \triangle_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ. Then $(x_0, (1, 0)) \notin WF(u)$.*

The proof of Theorem 1 makes use of the following lemma.

LEMMA 3. *Let u be in $\mathcal{S}'(\mathbf{R}^2)$. Let h be a function such that $\hat{h} \in \mathcal{D}(\mathbf{R}^2)$ and* supp $h \subset \Omega$, *where $\Omega = [c_1, c_2] \times [d_1, d_2]$, with $0 < c_1 < c_2$ and $d_1 < 0 < d_2$. Let ϕ be a function in $\mathcal{E}(\mathbf{R}^2)$, each of whose derivatives is polynomially bounded. Let Z be a compact set in $\mathbf{R}^2$ with $Z \cap$ supp $\phi = \emptyset$. Let Γ be an open cone symmetric with respect to $(1, 0)$ with opening half-angle less than $\frac{\pi}{2}$. Then $(L_h \phi u)|_{(Z, \triangle_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ.*

See the appendix for the proof.

Proof of Theorem 1. Suppose that $(x_0, (1, 0)) \notin WF(u)$. Then $\widehat{\psi u}$ decreases rapidly in some open conic neighborhood Γ_1 around $(1, 0)$ for some ψ in $\mathcal{D}(\mathbf{R}^2)$ with $\psi = 1$ in a neighborhood of x_0.

Let $\xi = (\xi_1, \xi_2)$ be in $\mathbf{R}^+ \times \mathbf{R}$. Then by Lemma 2, there is a nonempty open cone $\Gamma \subset \Gamma_1$ around $(1,0)$ such that if $\xi \in \Gamma$ with $|\xi|$ sufficiently large,

$$O_\xi \equiv \bigcup\nolimits_{(a,\tau)\in\triangle_\xi} \mathcal{V}(a,\tau)\Omega \subset \Psi_\xi,$$

where

$$\Psi_\xi \equiv \left[\xi_1\frac{c_1}{c_2}, \xi_1\frac{c_2}{c_1}\right] \times \left[\xi_2 - (d_2-d_1)\sqrt{\frac{\xi_1}{c_1}},\ \xi_2 + (d_2-d_1)\sqrt{\frac{\xi_1}{c_1}}\right],$$

and $\Psi_\xi \subset \Gamma_1$ for all (a,τ) in $\triangle_\xi$.

Now, note that

$$\begin{aligned}(L_h u)(a,b,\tau,\zeta) &= u[\overline{\zeta E_\tau T_b J_a h}] \\ &= \psi u[\overline{\zeta E_\tau T_b J_a h}] + (1-\psi)u[\overline{\zeta E_\tau T_b J_a h}],\end{aligned}$$

and since $\psi u \in \mathcal{E}'(\mathbf{R}^2)$, it follows that

$$\begin{aligned}&(L_h \psi u)(a,b,\tau,\zeta) = \psi u[\overline{\zeta E_\tau T_b J_a h}] \\ &= \int_{\mathbf{R}^2} \widehat{\psi u}(\eta)\overline{(\zeta E_\tau T_b J_a h)\,\widehat{}\,(\eta)}\ d\eta \\ &= \int_{\mathbf{R}^2} \widehat{\psi u}(\eta)\overline{\zeta}\epsilon^{2\pi i b\cdot(\eta_1,\eta_2-\tau)}\sqrt{|\det M(a)|}\ \ \overline{\hat{h}(a^2\eta_1, a(\eta_2-\tau))}d\eta.\end{aligned}$$

The integrand is nonzero only if $(a^2\eta_1, a(\eta_2-\tau)) \in \operatorname{supp}\hat{h}$, which means that $(\eta_1,\eta_2) \in \mathcal{V}(a,\tau)\Omega$.

Thus for each b in $\mathbf{R}^2$, (a,τ) in $\triangle_\xi$, and ζ in Q,

$$\begin{aligned}&|(L_h\psi u)(a,b,\tau,\zeta)| \\ &\le \sqrt{|\det M(a)|}\ \|\hat{h}\|_\infty \int_{O_\xi} |\widehat{\psi u}(\eta)|d\eta.\end{aligned}$$

But $O_\xi \subset \Gamma_1$ for all ξ in Γ with $|\xi|$ sufficiently large, and since $\widehat{\psi u}$ decreases rapidly in Γ_1, for each $N = 1,2,\ldots$, there is a constant C_N such that

$$\begin{aligned}&|(L_h\psi u)(a,b,\tau,\zeta)| \le C_N\sqrt{|\det M(a)|}\ \|\hat{h}\|_\infty \int_{O_\xi} \frac{1}{(1+|\eta|)^N}d\eta \\ &\le \left|\xi_1\frac{c_2}{c_1}\right|^{\frac{3}{2}}\ \|\hat{h}\|_\infty\ C_N\frac{2(d_2-d_1)\sqrt{\frac{\xi_1}{c_1}}}{N-1}\ \frac{1}{(1+|\xi_1\frac{c_1}{c_2}|)^{N-1}} \\ &\to 0 \quad \text{as} \quad |\xi| \to \infty\end{aligned}$$

for all ξ in Γ, where N may be chosen arbitrarily large.

On the other hand, $(1-\psi)\,u\,[\overline{\zeta E_\tau T_b J_a h}] = (L_h(1-\psi)u)(a,b,\tau,\zeta)$.

Let X be an open neighborhood of x_0 with $\overline{X} \cap \operatorname{supp}(1-\psi) = \emptyset$. Since $(1-\psi)$ has bounded derivatives and since Γ_1 is an open cone around $(1,0)$, it follows by Lemma 3 that $L_h((1-\psi)u)|_{(X,\triangle_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ_1.

Thus $(L_h u)|_{(X,\triangle_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ.

This proves the first part of Theorem 1.

In order to prove the second part of Theorem 1, we will apply the following theorem by Grossmann, Morlet, and Paul.

THEOREM 2 (see [2]). *Let U be strongly continuous square integrable unitary representation of a locally compact topological group G, acting on the Hilbert space H. Then there exists in H a unique self-adjoint positive operator C such that the following hold:*

(i) *The set of admissible vectors coincides with the domain of C, where the admissibility condition for h means $\int_G |\langle h, U(\gamma)h\rangle|^2 d_\gamma < \infty$, with γ in G, and where d_γ is the left Haar measure on G.*

(ii) *For f, g, h in H with h admissible,*

$$\int_G \langle f, U(\gamma)h\rangle \, \overline{\langle g, U(\gamma)h\rangle} d\gamma = C_h \langle f, g\rangle,$$

where $C_h = \langle Ch, Ch\rangle$.

See [2, p. 2475] for the proof.

Remark 3. U is said to be square-integrable if U is irreducible and there is in H at least one nonzero admissible vector [3].

Now, by considering the 4-parameter family of operators $U(a, b, \tau, \zeta) = \zeta E_\tau T_b J_a$, where $(a, b, \tau, \zeta) \in G$, then U is a strongly continuous unitary irreducible representation of the group G acting on the Hilbert space $L^2(\mathbf{R}^2)$. Next, suppose that h satisfies the hypotheses of Theorem 1. Then, in particular, $\hat{h}(0, k_2) = 0$ for all k_2 in $\mathbf{R}$. This implies that

$$C_h \equiv \int_{\mathbf{R}} \int_{\mathbf{R}} |\hat{h}(k_1, k_2)|^2 \frac{1}{2|k_1|} dk_1 \, dk_2 < \infty.$$

But it can be shown that

$$\int_{\mathbf{R}} \int_{\mathbf{R}} |\hat{h}(k_1, k_2)|^2 \frac{1}{2|k_1|} dk_1 \, dk_2 < \infty$$

if and only if

$$\int_G |\langle h, U(a, b, \tau, \zeta)h\rangle|^2 d(a, b, \tau, \zeta) < \infty.$$

Thus h is admissible. Hence, by Theorem 2, for f, g in $L^2(\mathbf{R}^2)$,

$$\int_G \langle f, U(a, b, \tau, \zeta)h\rangle \, \overline{\langle g, U(a, b, \tau, \zeta)h\rangle} d(a, b, \tau, \zeta) = C_h \langle f, g\rangle.$$

Let us now give the proof of the second part of Theorem 1.

Suppose that there is a function ψ in $\mathcal{D}(\mathbf{R}^2)$ with $\psi = 1$ in a neighborhood of x_0, an open neighborhood X of supp ψ, and an open conic neighborhood Γ of (1, 0) such that $(L_h \psi u)|_{(X, \Delta_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ. Without loss of generality, it will be assumed that X is bounded.

Let $Y = \text{supp } \psi$. Let A be a closed ball centered at x_0 of radius $\mathbf{R} > 0$ such that $X \subset A$ and set $L = \text{dist}(\partial Y, \partial X)$. Note that $L > 0$ by hypothesis. Let $v = \psi u$. Then $v \in \mathcal{E}'(\mathbf{R}^2)$ and supp $v \subset Y$.

Then the following two claims are needed.

CLAIM 1. *Let $\widehat{w}(\xi) = \int_G (L_h v)(a, b, \tau, \zeta)(U(a, b, \tau, \zeta)h)\hat{}(\xi) d(a, b, \tau, \zeta)$. Then $\widehat{w}(\xi)$ is well defined for all ξ in the cone Γ, and $\widehat{w}$ is rapidly decreasing in Γ.*

CLAIM 2. *For any function ϕ with $\breve{\phi} \in \mathcal{D}(\Gamma)$,*

$$\frac{1}{C_h}\int_\Gamma \hat{w}(\xi)\breve{\phi}(\xi)\,d\xi = v[\phi].$$

Finally, note that by Claim 2, $\widehat{v}|_\Gamma = \frac{1}{C_h}\,\widehat{w}$, and by Claim 1, $\frac{1}{C_h}\,\widehat{w}$ decreases rapidly in Γ. Therefore, $(x_0,(1,0)) \notin WF(u)$.

This completes the proof of Theorem 1.

Proof of Claim 1. Let ξ be in Γ. Then

$$|\hat{w}\,(\xi)| \le \int_0^1 \int_{\mathbf{R^2}} \int_{\mathbf{R}} \int_{\mathbf{R}} |(L_h v)(a,b,\tau,\zeta)||\ \hat{h}(a^2\xi_1, a(\xi_2\,-\,\tau))|\frac{1}{\sqrt{|\det M(a)|}}\,da\,d\tau\,db\,dr.$$

Note that $\hat{h}(a^2\xi_1, a(\xi_2-\tau)) \neq 0$ only if $(a,\,\tau) \in\, \Delta_\xi$. Then

$$|\hat{w}(\xi)| \le ||\hat{h}||_\infty\,(I_1(\xi)\,+\,I_2(\xi)\,+I_3(\xi)),$$

where

$$I_1(\xi) = \int_0^1 \int_X \int_{\Delta_\xi} |(L_h v)(a,\,b,\,\tau,\,\zeta)|\frac{1}{\sqrt{|\det M(a)|}}\,da\,d\tau\,db\,dr,$$

$$I_2(\xi) = \int_0^1 \int_{A\setminus X} \int_{\Delta_\xi} |(L_h v)(a,\,b,\,\tau,\,\zeta)|\frac{1}{\sqrt{|\det M(a)|}}\ \ da\,d\tau\,db\,dr,$$

and

$$I_3(\xi) = \int_0^1 \int_{\mathbf{R}^2\setminus A} \int_{\Delta_\xi} |(L_h v)(a,\,b,\,\tau,\,\zeta)|\frac{1}{\sqrt{|\det M(a)|}}\ \ da\,d\tau\,db\,dr.$$

Now, by hypothesis, $(L_h v)|_{(X,\Delta_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ for all ξ in Γ. Then for each $N = 1,\,2,\ldots,$ there is a constant P_N such that for all ξ in Γ,

$$\begin{aligned} I_1(\xi) &\le \frac{P_N}{(1\,+\,|\xi|)^N}\ \ m(X)\int_{\Delta_\xi}\frac{1}{|a|^{\frac{3}{2}}}da\,d\tau \\ &\le \frac{P_N}{(1\,+\,|\xi|)^N}\ \ m(X)(d_2-d_1)\left(\sqrt{\frac{\xi_1}{c_1}}\right)^{\frac{5}{2}}\left(\sqrt{\frac{c_2}{\xi_1}}-\sqrt{\frac{c_1}{\xi_1}}\right) \\ &\to 0 \quad \text{as} \quad |\xi|\to\infty \quad \text{in} \quad \Gamma, \end{aligned}$$

where N may be chosen arbitrarily large.

Consider now

$$I_2(\xi) = \int_0^1 \int_{A\setminus X} \int_{\Delta_\xi} |(L_h v)(a,\,b,\,\tau,\,\zeta)|\frac{1}{\sqrt{|\det M(a)|}}\ \ da\,d\tau\,db\,dr.$$

Because $A\setminus X$ is a compact set and $(A\setminus X)\cap$ supp ψ is empty, and since

$$(L_h v)(a,\,b,\,\tau,\,\zeta) = u[\psi\overline{\zeta\,E_\tau T_b\,J_a\,h}],$$

it follows from Lemma 3 that $L_h(\psi u)|_{(A \setminus X, \Delta_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ.

Then for each $N = 1, 2, \ldots$, there is a constant R_N such that for all ξ in Γ,

$$I_2(\xi) \le \frac{R_N}{(1 + |\xi|)^N} m(A \setminus X)(d_2 - d_1) \left(\sqrt{\frac{\xi_1}{c_1}}\right)^{\frac{5}{2}} \left(\sqrt{\frac{c_2}{\xi_1}} - \sqrt{\frac{c_1}{\xi_1}}\right).$$

Thus, $I_2(\xi) \to 0$ rapidly as $|\xi| \to \infty$ in Γ.

Now for

$$I_3(\xi) = \int_0^1 \int_{\mathbf{R}^2 \setminus A} \int_{\Delta_\xi} |(L_h v)(a, b, \tau, \zeta)| \frac{1}{\sqrt{|\det M(a)|}} \, da \, d\tau \, db \, dr$$

note that, since $u \in \mathcal{S}'(\mathbf{R}^2)$ and supp $\psi = Y$, it follows that

$$(L_h v)(a, b, \tau, \zeta) = \int_Y (-1)^{|\alpha|} g(x) D_x^\alpha \psi \overline{\zeta E_\tau T_b J_a h(x)} \, dx$$

for some polynomially bounded continuous function g and some multiindex α.

Because $\psi \in \mathcal{D}(\mathbf{R}^2)$, there is a constant $M_1 > 0$ such that $|D^\delta \psi(x)| \le M_1$ for all multiindices δ with $|\delta| \le |\alpha|$ and all x in $\mathbf{R}^2$. Also, because $h \in \mathcal{S}(\mathbf{R}^2)$, for each $N = 1, 2, \ldots$, there is a constant $B_N > 0$ such that $|D^\gamma h(x)| \le \frac{B_N}{(1+|x|)^N}$ for all x in $\mathbf{R}^2$, and since g is a continuous function and Y is compact, it follows that there is a constant $M_2 > 0$ such that $|g(x)| \le M_2$ for all x in Y. Hence,

$$\begin{aligned} &|(L_h v)(a, b, \tau, \zeta)| \\ &\le M_1 M_2 B_N \sum_{\beta \le \alpha} \sum_{\substack{\gamma \le \beta \\ \gamma_1 = \beta_1}} \binom{\beta}{\gamma} \frac{|2\pi\tau|^{\beta_2 - \gamma_2}}{|a|^{\frac{3}{2}}} \int_Y \frac{1}{(1 + |M(a)^{-1}(x - b)|)^N} dx. \end{aligned}$$

Note that if $\xi \in \Gamma$ and $|\xi|$ is sufficiently large, then $|a| < 1$ for $(a, \tau) \in \Delta_\xi$.

In this case $\frac{1}{a^2} \le \frac{1}{a^4}$, and $\frac{1}{|a|}|x - b| \le |M(a)^{-1}(x - b)|$. Note also that for $x \in Y$ and $b \in \mathbf{R}^2 \setminus A$, we have $0 < L + (|x_0 - b| - R) \le |x - b|$. Then

$$\int_Y \frac{1}{(1 + \frac{1}{|a|}|x - b|)^N} dx \le m(Y) \frac{1}{(1 + \frac{1}{|a|}(L + |x_0 - b| - R))^N}.$$

Hence,

$$\begin{aligned} I_3(\xi) \le M_1 M_2 B_N \sum_{\alpha \le \beta} \sum_{\substack{\gamma \le \beta \\ \gamma_1 = \beta_1}} \binom{\beta}{\gamma} |2\pi|^{\beta_2 - \gamma_2} m(Y) \\ \cdot \int_{\mathbf{R}^2 \setminus A} \int_{\Delta_\xi} \frac{|\tau|^{\beta_2 - \gamma_2}}{|a|^3 (1 + \frac{L + |x_0 - b| - R}{|a|})^N} da \, d\tau \, db. \end{aligned}$$

Note that for (a, τ) in Δ_ξ, $\sqrt{c_1/\xi_1} \le |a| \le \sqrt{c_2/\xi_1}$ and $|\tau| \le |\xi_2| + (d_2 + |d_1|)\sqrt{\xi_1/c_1}$, and since $L + |x_0 - b| - R > 0$, it follows that

$$\frac{1}{1 + \frac{L + |x_0 - b| - R}{|a|}} \le \frac{1}{1 + (L + |x_0 - b| - R)\sqrt{\frac{\xi_1}{c_2}}}.$$

Hence,

$$\int_{\Delta_\xi} \frac{|\tau|^{\beta_2-\gamma_2}}{|a|^3\left(1+\frac{L+|x_0-b|-R}{|a|}\right)^N}da\,d\tau$$
$$\le (d_2-d_1)\left(\sqrt{\frac{c_2}{\xi_1}}-\sqrt{\frac{c_1}{\xi_1}}\right)\left(\frac{\xi_1}{c_1}\right)^2\left(|\xi_2|+(d_2+|d_1|)\sqrt{\frac{\xi_1}{c_1}}\right)^{\beta_2-\gamma_2}$$
$$\cdot\frac{1}{(1+(L+|x_0-b|-R)\sqrt{\frac{\xi_1}{c_2}})^N}.$$

But

$$\int_{\mathrm{R}^2\setminus A}\frac{1}{(1+(L+|x_0-b|-R)\sqrt{\frac{\xi_1}{c_2}})^N}db$$
$$=2\pi\left|\frac{c_2}{\xi_1}\right|\frac{1}{N-2}\frac{1}{\left(1+L\sqrt{\frac{\xi_1}{c_2}}\right)^{N-2}}$$
$$+2\pi\sqrt{\frac{c_2}{\xi_1}}\left(-\sqrt{\frac{c_2}{\xi_1}}-L+R\right)\frac{1}{N-1}\frac{1}{\left(1+L\sqrt{\frac{\xi_1}{c_2}}\right)^{N-1}}$$
$$\to 0\quad\text{as}\quad|\xi|\to\infty\quad\text{for } N \text{ sufficiently large.}$$

Thus, $\widehat{w}$ decreases rapidly in Γ. This completes the proof of Claim 1. The following lemma is used to prove Claim 2.

LEMMA 2. *Let v be in $\mathcal{E}'(\mathbf{R}^2)$. Then for h admissible in $\mathcal{S}(\mathbf{R}^2)$,*

$$(L_h v)(a,\,b,\,\tau,\,\zeta)=\sum_{\beta\le\alpha}\sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}}\binom{\beta}{\gamma}(-1)^{|\gamma|}(-2\pi i\tau)^{\beta_2-\gamma_2}D_b^\gamma(L_h g_\beta)(a,\,b,\,\tau,\,\zeta)$$

for some multiindex α and some compactly supported continuous functions g_β.

See the appendix for the proof.

Proof of Claim 2. Let ξ be in Γ. Then by Claim 1,

$$\hat{w}(\xi)=\int_G (L_h v)(a,\,b,\,\tau,\,\zeta)(U(a,\,b,\,\tau,\,\zeta)h)\widehat{\ }(\xi)d(a,\,b,\,\tau,\,\zeta)$$

is well defined for all ξ in Γ. Since $v=\psi u\in\mathcal{E}'(\mathbf{R}^2)$, it follows from Lemma 4 that

$$\hat{w}(\xi)=\int_G\sum_{\beta\le\alpha}\sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}}\binom{\beta}{\gamma}(-1)^{|\gamma|}(-2\pi i\tau)^{\beta_2-\gamma_2}D_b^\gamma(L_h g_\beta)(a,\,b,\,\tau,\,\zeta)$$
$$\cdot(U(a,\,b,\,\tau,\,\zeta)h)\widehat{\ }(\xi)d(a,\,b,\,\tau,\,\zeta)$$

for some multiindex α and some compactly supported continuous functions g_β. In particular, $g_\beta\in L^2(\mathbf{R}^2)$. Then for any function ϕ such that $\phi\in\mathcal{D}(\Gamma)$,

$$\int_\Gamma \hat{w}(\xi)\breve{\phi}(\xi)\,d\xi$$

$$
\begin{aligned}
&= \sum_{\beta\le\alpha}\int_G \sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}} \binom{\beta}{\gamma}(-1)^{|\gamma|}(-2\pi i\tau)^{\beta_2-\gamma_2}\ D_b^\gamma\ \langle g_\beta, U(a,b,\tau,\zeta)h\rangle \\
&\quad\cdot\left(\int_\Gamma (U(a,b,\tau,\zeta)h)\hat{}\,(\xi)\breve{\phi}(\xi)d\xi\right)\ d(a,b,\tau,\zeta) \\
&= \sum_{\beta\le\alpha}(-1)^{|\beta|}\int_G \sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}} \binom{\beta}{\gamma}(2\pi i\tau)^{\beta_2-\gamma_2} \\
&\quad\cdot\left(\int_\Gamma D_b^\gamma\langle g_\beta, U(a,b,\tau,\zeta)h\rangle(U(a,b,\tau,\zeta)h)(y)db\right)\phi(y)d(a,y,\tau,\zeta) \\
&= \sum_{\beta\le\alpha}(-1)^{|\beta|}\int_G \sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}} \binom{\beta}{\gamma}(2\pi i\tau)^{\beta_2-\gamma_2} \\
&\quad\cdot(-1)^{|\gamma|}\left(\int_{\mathbf{R}^2}\langle g_\beta, U(a,b,\tau,\zeta)h\rangle D_b^\gamma U(a,b,\tau,\zeta)h(y)db\right)\phi(y)d(a,y,\tau,\zeta) \\
&= \sum_{\beta\le\alpha}(-1)^{|\beta|}\int_G \langle g_\beta, U(a,b,\tau,\zeta)h\rangle \\
&\quad\cdot\left(\int_{\mathbf{R}^2}\sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}} \binom{\beta}{\gamma}(2\pi i\tau)^{\beta_2-\gamma_2}(-1)^{|\gamma|}D_b^\gamma U(a,b,\tau,\zeta)h(y)\phi(y)dy\right)d(a,b,\tau,\zeta) \\
&= \sum_{\beta\le\alpha}(-1)^{|\beta|}\int_G \langle g_\beta, U(a,b,\tau,\zeta)h\rangle \\
&\quad\cdot\left(\int_{\mathbf{R}^2} D_y^\beta U(a,b,\tau,\zeta)h(y)\phi(y)dy\right)d(a,b,\tau,\zeta) \\
&= \sum_{\beta\le\alpha}\int_G \langle g_\beta, U(a,b,\tau,\zeta)h\rangle\left(\int_{\mathbf{R}^2} U(a,b,\tau,\zeta)h(y)D_y^\beta\phi(y)dy\right)d(a,b,\tau,\zeta) \\
&= \sum_{\beta\le\alpha}\int_G \langle g_\beta, U(a,b,\tau,\zeta)h\rangle\overline{\langle\overline{D^\beta\phi},\ U(a,b,\tau,\zeta)h\rangle}d(a,b,\tau,\zeta) \\
&= \sum_{\beta\le\alpha} C_h\langle g_\beta, \overline{D^\beta\phi}\rangle \\
&= C_h v[\phi].
\end{aligned}
$$

This completes the proof of Claim 2.

Example. For $x=(x_1,x_2)$ in $\mathbf{R}^2$, let $u(x_1,x_2)=\delta(x_1)$. Then $u\in\mathcal{S}'(\mathbf{R}^2)$ and sing*supp* $u=\{(0,x_2): x_2\in\mathbf{R}\}$. Note that the action of u on a test function ϕ in $\mathcal{S}(\mathbf{R}^2)$ is given by

$$u[\phi]=\int_{-\infty}^{\infty}\phi(0,x_2)dx_2,$$

and the wave front set of u is

$$WF(u)=\{((0,x_2);(\xi_1,0)): x_2\in\mathbf{R}\ \text{ and }\ \xi_1\in\mathbf{R}\setminus\{0\}\}.$$

Let h be a function such that h is nonnegative, $\hat{h}\in\mathcal{D}(\mathbf{R}^2)$, $\hat{h}$ is not identically zero, and supp$\hat{h}\subset\Omega$, where $\Omega=[c_1,c_2]\times[d_1,d_2]$ with $0<c_1<c_2, d_1<0<d_2$.

Let $\xi = (\xi_1, \xi_2)$ be in $\mathbf{R}^+ \times \mathbf{R}$, and let Δ_ξ be as defined. For $(x_0, (1,0))$ in $\mathbf{R}^2 \times \mathbf{R}^2$, where $x_0 = (x_{01}, x_{02})$, let us consider two cases.

Case 1. Suppose that $x_{01} \neq 0$. Then $((x_{01}, x_{02}); (1,0)) \notin WF(u)$. So $\widehat{\psi u}$ decreases rapidly in some cone Γ around (1, 0) for some $\psi \in \mathcal{D}(\mathbf{R}^2)$ with $\psi = 1$ in a neighborhood X of $x_0 = (x_{01}, x_{02})$. Let us show that $(L_h u)|_{(X,\Delta_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ.

Let (a, τ) be in Δ_ξ, $b = (b_1, b_2)$ in X, and ζ in Q. Note that if $|\xi| \to \infty$ in Γ, then $a \to 0$. Also, if $b \in X$, then $b_1 > 0$.

Now,

$$(L_h u)(a,b,\tau,\zeta) = u[\overline{\zeta E_\tau T_b J_a h}] = \int_{-\infty}^{\infty} \overline{\zeta E_\tau T_b J_a h(0,x_2)} dx_2$$
$$= \int_{-\infty}^{\infty} \overline{\zeta} e^{-2\pi i \tau x_2} \frac{1}{|a|^{\frac{3}{2}}} \overline{h\left(-\frac{b_1}{a^2}, \frac{x_2 - b_2}{a}\right)} dx_2.$$

Then

$$|(L_h u)(a,b,\tau,\zeta)| \leq \frac{1}{|a|^{\frac{3}{2}}} \int_{-\infty}^{\infty} \left| h\left(-\frac{b_1}{a^2}, \frac{x_2 - b_2}{a}\right)\right| dx_2$$
$$= \frac{1}{|a|^{\frac{1}{2}}} \int_{-\infty}^{\infty} \left| h\left(-\frac{b_1}{a^2}, y\right)\right| dy.$$

Because $h \in \mathcal{S}(\mathbf{R}^2)$,

$$\left| h\left(-\frac{b_1}{a^2}, y\right)\right| \leq \frac{C_N}{\left(\sqrt{\frac{b_1^2}{a^4} + y^2}\right)^N}, \qquad N = 1, 2, \ldots.$$

Thus,

$$|(L_h u)(a,b,\tau,\zeta)| \leq C_N a^{2N} |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} \frac{1}{(\sqrt{b_1^2 + a^4 y^2})^N} dy$$
$$= C_N a^{2N} |a|^{-\frac{5}{2}} \int_{-\infty}^{\infty} \frac{1}{(b_1^2 + z^2)^{\frac{N}{2}}} dz$$
$$\to 0 \quad \text{as} \quad a \to 0,$$

where N can be chosen arbitrarily large.

Therefore, $(L_h u)|_{(X,\Delta_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ for all ξ in Γ.

Case 2. Suppose that $x_{01} = 0$. Then $((0, x_{02}); (1,0)) \in WF(u)$.

Let Γ be any open cone symmetric with respect to (1, 0). For $\xi = (\xi_1, 0)$ in Γ, Δ_ξ is the set of points (a, τ) in $(\mathbf{R} \setminus \{0\}) \times \mathbf{R}$ such that $\sqrt{c_1/\xi_1} \leq |a| \leq \sqrt{c_2/\xi_1}$ and $-\frac{d_2}{|a|} \leq \tau \leq -\frac{d_1}{|a|}$.

Let us show that $(L_h u)|_{(X,\Delta_\xi)}$ does not fall off rapidly as $|\xi| = \xi_1 \to \infty$, where X is any open neighborhood of $x_0 = (0, x_{02})$. Then for $b = (0, b_2)$ in X and $(a, 0)$ in $\Delta_{(\xi_1,0)}$,

$$(L_h u)(a,b,0,\zeta) = u[\overline{\zeta E_0 T_b J_a h}] = \int_{-\infty}^{\infty} \overline{\zeta E_0 T_b J_a h(0,x_2)} dx_2$$
$$= \int_{-\infty}^{\infty} \overline{\zeta} \frac{1}{|a|^{\frac{3}{2}}} \overline{h\left(0, \frac{x_2 - b_2}{a}\right)} dx_2 = \overline{\zeta} \frac{1}{\sqrt{|a|}} \int_{-\infty}^{\infty} \overline{h(0, z_2)} dz_2.$$

Let $A = \int_{-\infty}^{\infty} \overline{h(0, z_2)} dz_2$. Then A is independent of a and $0 < A < \infty$;

$$\sup_{\substack{b\in X\\(a,\tau)\in\Delta_\xi\\ \zeta\in Q}} |(L_h u)(a, b, \tau, \zeta)| \geq \frac{1}{\sqrt{|a|}} A \geq \left(\frac{\xi_1}{c_2}\right)^{\frac{1}{4}} A.$$

Thus, $(L_h u)|_{(X,\Delta_\xi)}$ does not fall off as $|\xi| = \xi_1 \to \infty$ in Γ.

Appendix.

Proof of Lemma 1. Let $\xi = (\xi_1, \xi_2)$ be in $\mathbf{R}^+ \times \mathbf{R}$. Then by Definition 4,

$$\begin{aligned}\Delta\xi &= \{(a,\tau) \in (\mathbf{R} \setminus \{0\}) \times \mathbf{R} : \ \mathcal{V}(a,\tau)\Omega \ni \xi\} \\ &= \{(a,\tau) \in (\mathbf{R} \setminus \{0\}) \times \mathbf{R} : \ \mathcal{V}^{-1}(a,\tau)\xi \in \Omega\},\end{aligned}$$

where by definition of $\mathcal{V}(a,\tau)$, $\mathcal{V}^{-1}(a,\tau)\xi = (a^2\xi_1, a(\xi_2 - \tau))$.

Then $\mathcal{V}^{-1}(a,\tau)\xi \in \Omega$ if $c_1 \leq a^2\xi_1 \leq c_2$ and $d_1 \leq a(\xi_2 - \tau) \leq d_2$; equivalently

$$\sqrt{c_1/\xi_1} \leq |a| \leq \sqrt{c_2/\xi_1} \ \text{ and } \ \xi_2 - \tfrac{d_2}{|a|} \leq \tau \leq \xi_2 - \tfrac{d_1}{|a|}.$$

This proves Lemma 1.

Proof of Lemma 2. Take $\alpha > 0$ such that $\tan\alpha < \frac{c_1}{c_2} \tan\beta$. Let $\epsilon = \tan\alpha$. Consider then $\Gamma = \{(\eta_1, \eta_2) \in \mathbf{R}^+ \times \mathbf{R} : |\eta_2| < \epsilon\eta_1\}$. Note that $\Gamma_1 \supset \{(\eta_1, \eta_2) \in \mathbf{R}^+ \times \mathbf{R} : |\eta_2| < \frac{c_2}{c_1}\epsilon\eta_1\}$ and since $0 < c_1 < c_2$, $\Gamma \subset \Gamma_1$.

Let $\xi = (\xi_1, \xi_2)$ be in Γ. Since $\Gamma \subset \mathbf{R}^+ \times \mathbf{R}$, it follows from Lemma 1 that

$$\Delta_\xi = \left\{(a,\tau) \in (\mathbf{R} \ \setminus \ \{0\}) \times \mathbf{R} : \sqrt{\frac{c_1}{\xi_1}} \leq |a| \leq \sqrt{\frac{c_2}{\xi_1}} \ \text{ and } \ \xi_2 - \frac{d_2}{|a|} \leq \tau \leq \xi_2 - \frac{d_1}{|a|}\right\}.$$

Now, let (η_1, η_2) be in $\mathcal{V}(a,\tau)\Omega$. Then by Remark 1, $\frac{c_1}{a^2} \leq \eta_1 \leq \frac{c_2}{a^2}$ and $\frac{d_1}{|a|} + \tau \leq \eta_2 \leq \frac{d_2}{|a|} + \tau$. Then for (a,τ) in Δ_ξ, $\frac{c_1}{c_2}\xi_1 \leq \eta_1 \leq \frac{c_2}{c_1}\xi_1$ and

$$\xi_2 - (d_2 - d_1)\sqrt{\tfrac{\xi_1}{c_1}} \leq \eta_2 \leq \xi_2 + (d_2 - d_1)\sqrt{\tfrac{\xi_1}{c_1}}.$$

That is, for (a,τ) in Δ_ξ, $\mathcal{V}(a,\tau)\Omega$ is contained in the rectangle

$$\Psi_\xi = \left[\tfrac{c_1}{c_2}\xi_1, \tfrac{c_2}{c_1}\xi_1\right] \times \left[\xi_2 - (d_2 - d_1)\sqrt{\tfrac{\xi_1}{c_1}}, \xi_2 + (d_2 - d_1)\sqrt{\tfrac{\xi_1}{c_1}}\right].$$

Suppose that $\xi_2 \geq 0$. Then $\xi_2 + (d_2 - d_1)\sqrt{\xi_1/c_1} > 0$. Let Θ be the upper left corner of the rectangle Ψ_ξ. That is, $\Theta = (\frac{c_1}{c_2}\xi_1, \xi_2 + (d_2 - d_1)\sqrt{\xi_1/c_1})$. Then the slope m_Θ of the line passing through (0, 0) and Θ is

$$m_\Theta = \frac{\xi_2 + (d_2 - d_1)\sqrt{\frac{\xi_1}{c_1}}}{\frac{c_1}{c_2}\xi_1} = \frac{c_2}{c_1}\frac{\xi_2}{\xi_1} + \frac{(d_2 - d_1)c_2}{{c_1}^{\frac{3}{2}}}\frac{1}{\sqrt{\xi_1}}.$$

Since $\xi \in \Gamma$, and $\xi_2 > 0$, it follows that $\frac{\xi_2}{\xi_1} < \epsilon$. Then

$$|\xi|^2 = \xi_1^2 + \xi_2^2 < \xi_1^2 + \epsilon^2\xi_1^2 = (1 + \epsilon^2)\xi_1^2; \ \frac{1}{\xi_1} < \frac{\sqrt{1+\epsilon^2}}{|\xi|}.$$

Thus,

$$m_\Theta < \frac{c_2}{c_1}\epsilon + \frac{(d_2-d_1)c_2}{{c_1}^{\frac{3}{2}}}\frac{(1+\epsilon^2)^{\frac{1}{4}}}{\sqrt{|\xi|}}.$$

Since $\tan\beta > \frac{c_2}{c_1}\epsilon$, it follows that for sufficiently large $|\xi|$, $m_\Theta < \tan\beta$. Thus $\Theta \in \Gamma_1$.

A similar analysis shows that the lower left corner of the rectangle Ψ_ξ is also in Γ_1. It follows that for sufficiently large $|\xi|$, $\mathcal{V}(a,\ \tau)\Omega \subset \Gamma_1$ for all $(a,\ \tau)$ in Δ_ξ.

This proves Lemma 2.

Proof of Lemma 3. The idea behind the proof is that, as $|\xi| \to \infty$ in Γ, the condition $(a,\ \tau) \in \Delta_\xi$ forces $a \to 0$, which concentrates the function $\zeta E_\tau T_b J_a h$ about the point $b \in Z$, where ϕ is zero.

Let $b \in Z, \xi \in \Gamma$, and let $(a,\tau) \in \Delta_\xi$. Since $u \in \mathcal{S}'(\mathbf{R}^2)$ and $\phi\overline{\zeta E_\tau T_b J_a h} \in \mathcal{S}(\mathbf{R}^2)$, it follows that

$$\begin{aligned}(\,L_{\,h}\phi u)(a,b,\tau,\zeta) &= \phi u[\overline{\zeta E_\tau T_b J_a h}]\\ &= u[\phi\overline{\zeta E_\tau T_b J_a h}] = \int_{\mathbf{R}^2}(-1)^{|\alpha|}g(x)D_x^a\,\phi\overline{\zeta E_\tau T_b J_a h}(x)dx\end{aligned}$$

for some polynomially bounded continuous function g and some multiindex α.

Because g is a polynomially bounded continuous function, $h \in \mathcal{S}(\mathbf{R}^2)$, each of the derivatives of ϕ is polynomially bounded, and $b \in Z$, there are constants K_1, K_2, and $L > 0$ where $B_L(b) \cap \operatorname{supp}\phi = \emptyset$, with L independent of b, such that for $N = 1, 2, \ldots$ there is a constant $D_N > 0$ such that

$$\begin{aligned}&|(L_h\phi u)(a,b,\tau,\zeta)|\\ &\le K_1 D_N \sum_{\beta\le\alpha}\binom{\alpha}{\beta}\sum_{\substack{\gamma\le\beta\\ \gamma_1=\beta_1}}\binom{\beta}{\gamma}\frac{|2\pi\tau|^{\beta_2-\gamma_2}}{|a|^{\frac{3}{2}}}\int_{\mathbf{R}^2\setminus B_L(b)}\frac{(1+|x|)^{K_2}}{(1+|M(a)^{-1}(x-b)|)^N}dx.\end{aligned}$$

Now note that if $\xi \in \Gamma$ and $|\xi|$ is sufficiently large, then $|a| < 1$ for $(a,\tau) \in \triangle_\xi$. In this case $\frac{1}{|a|}|x-b| \le |M(a)^{-1}(x-b)|$. Then

$$\begin{aligned}&\int_{\mathbf{R}^2\setminus B_L(b)}\frac{(1+|x|)^{K_2}}{(1+|M(a)^{-1}(x-b)|)^N}dx\\ &\le 2\pi|a|\sum_{\rho=0}^{K_2}\binom{K_2}{\rho}|b|^{K_2-\rho}\frac{1}{N-2-\rho}\quad\frac{1}{\left(1+\frac{L}{|a|}\right)^{N-2-\rho}}.\end{aligned}$$

Since for (a,τ) in $\triangle_\xi$,

$$\sqrt{\frac{c_1}{\xi_1}} \le |a| \le \sqrt{\frac{c_2}{\xi_1}}\quad\text{and}\quad|\tau| \le |\xi_2| + (d_2+|d_1|)\sqrt{\frac{\xi_1}{c_1}};$$

it follows that

$$|(L_h\phi u)(a,b,\tau,\zeta)|$$

$$\leq K_1 D_N \sum_{\beta\leq\alpha} \binom{\alpha}{\beta} \sum_{\substack{\gamma\leq\beta \\ \gamma_1=\beta_1}} \binom{\beta}{\gamma} |2\pi|^{\beta_2-\gamma_2+1} \left(|\xi_2| + (d_2+|d_1|)\sqrt{\frac{\xi_1}{c_1}}\right)^{\beta_2-\gamma_2}$$

$$\cdot \left(\sqrt{\frac{\xi_1}{c_1}}\right)^{\frac{1}{2}} \sum_{\rho=0}^{K_2} \binom{K_2}{\rho} |b|^{K_2-\rho} \frac{1}{N-2-\rho} \quad \frac{1}{\left(1+L\sqrt{\frac{\xi_1}{c_2}}\right)^{N-2-\rho}}$$

$$\to 0 \text{ as } |\xi| \to \infty \text{ in } \Gamma.$$

Because N may be chosen arbitrarily large, the decrease is rapid, and since L is independent of $b \in Z$, it follows that $(L_h \phi u)|_{(Z,\triangle_\xi)} \to 0$ rapidly as $|\xi| \to \infty$ in Γ.

This proves Lemma 3.

Proof of Lemma 4. Since $u \in \mathcal{E}'(\mathbf{R}^2)$ and $h \in \mathcal{S}(\mathbf{R}^2)$, it follows that

$$u[\zeta E_\tau T_b J_a h] = \sum_{\beta\leq\alpha} \int_{\mathbf{R}^2} g_\beta(x) D_x^\beta \zeta E_\tau T_b J_a h(x) dx$$

for some multiindex α and compactly supported continuous functions g_β. Then

$$\begin{aligned}
(L_h u)(a,b,\tau,\zeta) &= u[\overline{\zeta E_\tau T_b J_a h}] \\
&= \sum_{\beta\leq\alpha} \int_{\mathbf{R}^2} g_\beta(x) D_x^\beta \overline{\zeta E_\tau T_b J_a h(x)} dx \\
&= \sum_{\beta\leq\alpha} \langle g_\beta, D_x^\beta \zeta E_\tau T_b J_a h\rangle \\
&= \sum_{\beta\leq\alpha} \sum_{\substack{\gamma\leq\beta \\ \gamma_1=\beta_1}} \binom{\beta}{\gamma} (-1)^{|\gamma|} (-2\pi i\tau)^{\beta_2-\gamma_2} D_b^\gamma \langle g_\beta, \zeta E_\tau T_b J_a h\rangle \\
&= \sum_{\beta\leq\alpha} \sum_{\substack{\gamma\leq\beta \\ \gamma_1=\beta_1}} \binom{\beta}{\gamma} (-1)^{|\gamma|} (-2\pi i\tau)^{\beta_2-\gamma_2} D_b^\gamma (L_h g_\beta)(a,b,\tau,\zeta).
\end{aligned}$$

This proves Lemma 4.

REFERENCES

[1] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

[2] A. Grossmann, J. Morlet, and T. Paul, *Transforms associated to square integrable group representations,* I. *General results*, J. Math. Phys., 26 (1985), pp. 2473–2479.

[3] A. Grossmann, J. Morlet and T. Paul, *Transforms associated to square integrable group representations,* II. *Examples*, Ann. Inst. H. Poincaré, 45 (1986), pp. 293–309.

[4] C. E. Heil and D. F. Walnut, *Continuous and discrete wavelet transforms*, SIAM Rev., 31 (1989), pp. 628–666.

[5] L. Hormander, *The Analysis of Linear Partial Differential Operators, Vol.* I–IV, Springer-Verlag, New York, 1983.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 469–478

A LYAPUNOV FUNCTION FOR TRIDIAGONAL COMPETITIVE-COOPERATIVE SYSTEMS*

BERNOLD FIEDLER† AND TOMÁŠ GEDEON‡

Abstract. We construct a Lyapunov function for tridiagonal competitive-cooperative systems. The same function is a Lyapunov function for Kolmogorov tridiagonal systems, which are defined on a closed positive orthant in $\mathbf{R}^n$. We show that all bounded orbits converge to the set of equilibria. Moreover, we show that there can be no heteroclinic cycles on the boundary of the first orthant, extending the results of H. I. Freedman and H. L. Smith [*Differential Equations Dynam. Systems*, 3 (1995), pp. 367–382].

Key words. Kolmogorov tridiagonal systems, Lyapunov function, heteroclinic cycles

AMS subject classifications. 34C37, 58F25

PII. S0036141097316147

1. Results.

We consider a system of differential equations

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2), \\ \dot{x}_i &= f_i(x_{i-1}, x_i, x_{i+1}), \qquad i = 2, \dots, n-1, \\ \dot{x}_n &= f_n(x_{n-1}, x_n), \end{aligned} \tag{1}$$

where functions f_i are defined on a nonempty open subset A of R^n. We assume that the f_i and their partial derivatives are continuous on A. We also assume that there are $\delta_i \in \{-1, +1\}$, such that

$$\delta_i \frac{\partial f_i}{\partial x_{i+1}} > 0, \quad \delta_i \frac{\partial f_{i+1}}{\partial x_i} > 0, \quad 1 \le i \le n-1.$$

This assumption implies that the Jacobi matrix $\partial f/\partial x$, corresponding to (1), is tridiagonal and sign symmetric in the sense that $\partial f_i/\partial x_{i+1}$ and $\partial f_{i+1}/\partial x_i$ have the same sign δ_i. If $\delta_i = -1$ for all i, then (1) is called *competitive*. If $\delta_i = 1$ for all i, then (1) is called *cooperative*. We introduce new variables, following Smith [S1]. We let $\bar{x}_i = \mu_i x_i$, $\mu_i \in \{\pm 1\}$, $1 \le i \le n$, with $\mu_1 = 1$, $\mu_i = \delta_{i-1}\mu_{i-1}$. Then the system (1) transforms into a new system of the same type with new

$$\bar{\delta}_i = \mu_i \mu_{i+1} \delta_i = \mu_i^2 \delta_i^2 = 1.$$

Therefore we can always assume, without loss of generality, that the competitive-cooperative system (1) is in fact cooperative and

$$\frac{\partial f_i}{\partial x_{i+1}} > 0, \quad \frac{\partial f_{i+1}}{\partial x_i} > 0, \quad 1 \le i \le n-1. \tag{H1}$$

*Received by the editors February 5, 1997; accepted for publication (in revised form) July 22, 1998; published electronically March 19, 1999. This research was partially supported by MONTS grants 219627 and 291725 and NSF grant DMS–291222.

http://www.siam.org/journals/sima/30-3/31614.html

†Institut für Mathematik I, Freie Universtät Berlin Arnimallee 2-6, D-14195 Berlin, Germany (fiedler@math.fu-berlin.de).

‡Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717-0240 (gedeon@poincare.math.montana.edu).

Our goal in this paper is to construct a Lyapunov function for a class of equations (1). A Lyapunov function $V : \mathbf{R}^n \to \mathbf{R}$ for (1) is a real valued function which is nonincreasing along trajectories of (1) and is strictly decreasing along all nonequilibrium trajectories. We assume that the system (2) is *dissipative*. We spell out the precise form of the assumption below. We remark that this assumption implies that all trajectories of (2) eventually enter a compact region of phase space. By LaSalle's invariance principle (see Hale and Koçak [HK]) the existence of a Lyapunov function then implies that each trajectory converges to the set of equilibria. Our dissipativeness condition takes the form

$$f_i(x_{i-1}, x_i, x_{i+1})x_i < 0 \quad \text{for} \quad |x_i| \geq C, |x_{i\pm1}| \leq |x_i| \tag{H2}$$

and some large constant C.

It is easy to see that this assumption forces any trajectory to enter the box $\{x \in \mathbf{R}^n \mid |x_i| \leq C \text{ for all } i\}$ at some finite time and then remain there.

Our final assumption on the functions f_i is more technical. Observe that assumption (H1) implies that for any fixed x_i the set $z(f_i, x_i)$ of points (x_{i-1}, x_{i+1}) satisfying $f_i(x_{i-1}, x_i, x_{i+1}) = 0$ is a curve in the (x_{i-1}, x_{i+1})-plane. Furthermore $z(f_i, x_i)$ is monotone with respect to both x_{i-1} and x_{i+1}. The following assumption implies that for all x_i the curve $z(f_i, x_i)$ is unbounded in both x_{i-1} and x_{i+1} directions. Assume

$$\lim_{x_k \to \infty} f_i(x_{i-1}, x_i, x_{i+1}) > 0, \quad \lim_{x_k \to -\infty} f_i(x_{i-1}, x_i, x_{i+1}) < 0 \tag{H3}$$

for both $k = i - 1$ and $k = i + 1$ and all x_i, x_{i-1} and x_{i+1}.

We state the main theorem of this paper.

THEOREM 1.1. *The system* (1) *with assumptions* (H1), (H2), (H3) *admits a Lyapunov function.*

The system (1) with assumption (H1) is a *monotone dynamical system*. There is an extensive literature on monotone dynamical systems, starting with the work of Hirsch [Hi1, Hi2, Hi3, Hi4] for monotone semiflows. The results of Hirsch and later improvements by Matano [M], Smith and Thieme [ST1, ST2], and Poláčik [P] established that most orbits of a strongly order-preserving semiflow converge to the set of equilibria. For references on the theory of monotone semiflows see the recent monograph by Smith [S2].

For the system (1) more is known: Smilie [Sm] has shown that all trajectories converge to the set of equilibria. He used an integer-valued Lyapunov function (nodal properties) to prove his result. The main consequence of the existence of real valued Lyapunov function V for the system (1), namely that all trajectories of (1) converge to the set of equilibria, is not new and was proved by Smilie [Sm].

The importance of the existence of the Lyapunov function V is that it can be used in a more general setting. We now consider the class of *Kolmogorov* systems, which model an interaction of populations, where every population interacts only with "neighboring" populations. They have the form

$$\begin{aligned} \dot{x}_1 &= x_1 f_1(x_1, x_2), \\ \dot{x}_i &= x_i f_i(x_{i-1}, x_i, x_{i+1}), \\ \dot{x}_n &= x_n f_n(x_{n-1}, x_n) \end{aligned} \tag{2}$$

for $i = 2, \ldots, n-1$. We assume that the functions f_i satisfy (H1), (H2), and (H3). Since x_i's represent population densities we restrict ourselves to the closed positive orthant

$$\mathcal{O} := \{x \in \mathbf{R}^n \mid x_i \geq 0 \text{ for } i = 1, \ldots, n\}.$$

For specific biological systems modeled by (2), see Freedman and Smith [F-S] and the references herein.

The results of Smilie do not apply to the system (2) directly. The crucial assumption (H1) for the right-hand side of (2) holds only in the interior of the positive orthant. We also notice that $x_j(0) = 0$ implies $x_j(t) = 0$ for all t. Therefore every boundary hyperplane of the first orthant is invariant under (2). Furthermore, if we restrict the set of equations to such a boundary hyperplane, we obtain a decoupled system of the same type as (2). Consequently, all faces and subfaces of the boundary of the positive orthant are invariant under (2); the restriction of the system (2) to such a set is of the same type as the system (2) itself. The assumption (H1) is satisfied only in the interior of the positive orthant and by the previous argument in the interior of every face and subface of the boundary of the positive orthant.

The question arises whether all trajectories still converge to the set of equilibria in this case. The result is obviously true if the trajectory stays bounded away from the boundary of the region where it starts, which may be the positive orthant $\mathcal{O}$ itself, or a face or subface of the boundary. But in the context of population dynamics, the trajectories approaching the boundary of a given region are important, since they represent a situation whereby a certain population goes extinct.

This problem was studied by Freedman and Smith [F-S]. Under some nondegeneracy assumptions they were able to show that every bounded orbit converges either to an equilibrium, or to a cycle of equilibria on the boundary of $\mathcal{O}$. A *cycle of equilibria* is a nonempty finite set of equilibria $\{E_1, \ldots, E_n\}$ such that

$$E_1 \to E_2 \to \cdots \to E_n \to E_1.$$

Here we write $E_1 \to E_2$ if E_1 and E_2 are equilibria of (2), not necessarily distinct, such that there exists a solution $x \in \mathcal{O}$ with $\lim_{t\to-\infty} x(t) = E_1$ and $\lim_{t\to\infty} x(t) = E_2$. For illustration of such a boundary cycle of equilibria see Figure 1.

In every boundary component the flow is convergent to the set of equilibria, but the boundary components are assembled together in such a way that they produce a cycle of equilibria.

The Lyapunov function V constructed in Theorem 1.1 turns out to be a Lyapunov function for system (2). Furthermore, V is defined, continuous on the boundary of $\mathcal{O}$, and restricts to the Lyapunov function on every face and subface of the boundary of $\mathcal{O}$.

THEOREM 1.2. *Consider system* (2) *on the closed positive orthant* $\mathcal{O}$. *Assume that the functions* f_i *satisfy the assumptions* (H1), (H2), *and* (H3). *Then there is a Lyapunov function* $V : \mathcal{O} \to \mathbf{R}$, *which is strictly decreasing outside the set of equilibria.*

We can now rule out the existence of the cycles of equilibria on the boundary of $\mathcal{O}$.

COROLLARY 1.3. *Consider system* (2) *on the closed positive orthant* $\mathcal{O}$. *Assume that the functions* f_i *satisfy the assumptions* (H1), (H2), *and* (H3). *Then every bounded trajectory with initial data in the closed positive orthant* $\mathcal{O}$ *converges to the set of equilibria. If the set of equilibria is finite, every trajectory converges to a single equilibrium. Moreover, cycles of equilibria in* $\mathcal{O}$ *do not exist.*

The cycles of equilibria do not exist in the large class of Lotka–Voltera systems. We now briefly describe results of Fiedler and Gedeon [FG].

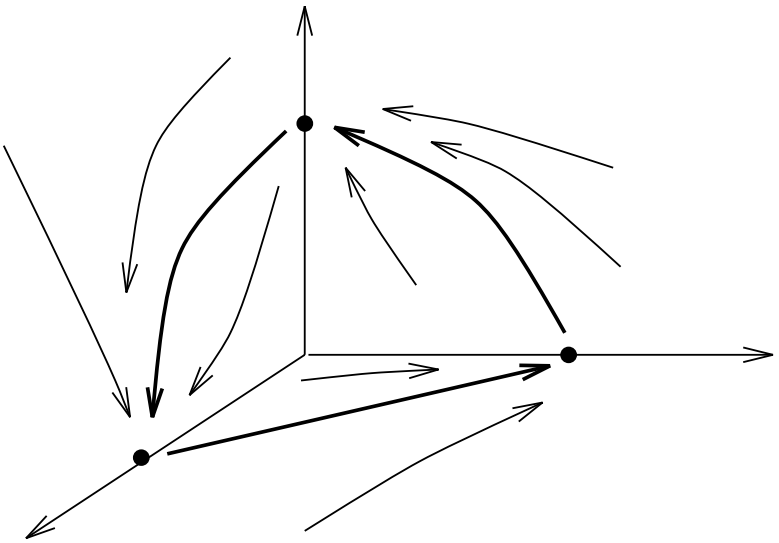


FIG. 1. *Heteroclinic cycle on the boundary of the positive orthant.*

Consider the system

$$\dot{x}_i = x_i \left(c_i - \sum_{j=1}^{n} \beta_{ij} f_j(x_j) \right) \tag{3}$$

on the positive orthant $\mathcal{O}$ in $\mathbf{R}^n$. If x_i describes the population size of a certain species, then the constants β_{ij} describe the interaction between the species. Let Υ be the undirected graph with n vertices, where the edge j is connected to the vertex i by the edge e_{ij} if and only if $\beta_{ij} \neq 0$. We assume $\beta_{ij}\beta_{ji} > 0$ for every edge e_{ij}. Therefore the definition of Υ makes sense.

Consider the system (3) and assume that $\frac{df_j}{dx_j} > 0$ for all j, that the interaction graph Υ is a tree, and that $\beta_{ij}\beta_{ji} > 0$ for every edge e_{ij}. Then, by the results of Fiedler and Gedeon [FG], every bounded trajectory of system (3) converges to the set of equilibria and boundary cycles of equilibria do not exist.

2. Proofs.

Proof of Theorem 1.1. Motivated by a somewhat analogous result of Matano [M] concerning parabolic partial differential equations with gradient dependence, we seek a Lyapunov function $V : \mathbf{R}^n \to \mathbf{R}$ of the form $V = -\sum_{i=1}^{n-1} g_i(x_i, x_{i+1})$. By differentiating and collecting terms we get

$$\dot{V} = -\sum_{i=1}^{n} \dot{x}_i(\partial_i g_i + \partial_i g_{i-1}),$$

where we used the notation $\partial_i := \frac{\partial}{\partial x_i}$. We also note that $\partial_1 g_0 = 0$ and $\partial_n g_n = 0$ in the expression for $\dot{V}$. Below, we will construct functions a_i such that

(a) $a_i f_i = \partial_i g_i + \partial_i g_{i-1}$ and $a_i > 0$ for $i = 1, \ldots, n$,

(b) $\partial_{i-1}(a_i f_i) = \partial_i(a_{i-1} f_{i-1})$ for $i = 2, \ldots, n$.

It follows from (a) that $a_i = a_i(x_{i-1}, x_i, x_{i+1})$ are functions of x_{i-1}, x_i, x_{i+1} for $i = 2, \ldots, n-1$, $a_1 = a_1(x_1, x_2)$, and $a_n = a_n(x_{n-1}, x_n)$. Property (a) also implies that

$$\begin{aligned}\dot{V} &= -\sum_i \dot{x}_i(\partial_i g_i + \partial_i g_{i-1}) \\ &= -\sum_i \dot{x}_i a_i f_i \\ &= -\sum_i a_i(\dot{x}_i)^2 \leq 0,\end{aligned}$$

since $a_i > 0$ for all i. Also $\dot{V}(x(t)) = 0$ if and only if $\dot{x}_i(t) = 0$ for all i. This implies that $\dot{V}(x(t)) = 0$ if and only if $x(t)$ is an equilibrium. Therefore V is a Lyapunov function.

We see that we have to construct functions a_i, $i = 1, \ldots, n$, which satisfy property (a).

Condition (b) is a consequence of (a) and the form of the function V. Indeed, (a) implies that

$$\partial_i(a_{i-1}f_{i-1}) = \partial_i\partial_{i-1}g_{i-1} + \partial_i\partial_{i-1}g_{i-2} = \partial_i\partial_{i-1}g_{i-1}$$

since $g_{i-2}(x_{i-2}, x_{i-1})$ does not depend on x_i. A similar computation leads to

$$\partial_{i-1}(a_i f_i) = \partial_i\partial_{i-1}g_{i-1}.$$

Therefore, if (a) is satisfied and g_i is a function of x_i and x_{i+1} only, then condition (b) must hold. In the inductive argument below we shall first construct functions $a_i > 0$ which satisfy (b), and then use (b) to construct functions g_i with property (a).

We construct the functions a_i, $i = 1, \ldots n$, by induction. The first step of the induction will be to construct functions a_1, g_0, and g_1 such that (a) and (b) are satisfied. We set $g_0 \equiv 0$ and $a_1 \equiv 1$. To determine $g_1(x_1, x_2)$ we set

$$\partial_1 g_1 := a_1 f_1 = f_1(x_1, x_2),$$

or more explicitly,

$$g_1(x_1, x_2) = \int_0^{x_1} f_1(\zeta_1, x_2) d\zeta_1.$$

This choice of a_1, g_0, and g_1 satisfies conditions (a) above. Observe that $\partial_2\partial_1 g_1(x_1, x_2)$ exists and is continuous. Condition (b) is vacuous for $i = 1$.

Having defined a_1, the condition (b) with $i = 2$ poses a restriction on a_2. The function a_2 must satisfy

$$\partial_1(a_2 f_2) = \partial_2(a_1 f_1).$$

Similarly, once the function a_{i-1} has been defined, the function a_i to be constructed in the next step of the induction must satisfy

$$\partial_{i-1}(a_i f_i) = \partial_i(a_{i-1} f_{i-1}).$$

With this in mind we define an auxiliary function

$$\gamma_1(x_1, x_2) := \partial_2(a_1 f_1).$$

Observe that $\partial_2(a_1 f_1) = \partial_2 f_1 > 0$, by assumption on function f_1, and so $\gamma_1 = \gamma_1(x_1, x_2) > 0$. So far, we have defined functions a_1, g_0, g_1 which satisfy (a), (b) for $i = 1$ and γ_1.

Now we proceed with the induction step. For technical reasons our induction hypothesis will not be statements (a) and (b) above but a slightly more complicated set of assumptions. We assume that we have constructed functions a_k, g_k, and γ_k, where $\gamma_k = \partial_{k+1}(a_k f_k)$ for $k = 1, \dots, i-1$ with the following properties:

(A1) $a_k > 0$ are continuous; $\partial_{k-1}(a_k f_k)$ and $\partial_{k+1}(a_k f_k)$ exist and are continuous.

(A2) $\partial_{k-1}(a_k f_k) = \partial_k(a_{k-1} f_{k-1})$.

(B) $a_k f_k = \partial_k g_k + \partial_k g_{k-1}$, $g_k = g_k(x_k, x_{k+1})$, and $\partial_{k+1}\partial_k g_k$ exists and is continuous.

(C) $\gamma_k(x_k, x_{k+1}) > 0$.

Observe that these conditions are satisfied for $k = 1$ by the above construction. Observe that (A2) is equivalent to (b), and (A1), (B) are equivalent to (a). Condition (C) is needed in the induction process.

Before we proceed with the induction step we introduce some notation. We denote $z(f_i) := \{x \mid f_i(x) = 0\}$ the zero set of the function f_i. In $\mathbf{R}^3$, spanned by coordinate axis x_{i-1}, x_i, x_{i+1}, the zero set $z(f_i)$ is a graph over the (x_{i-1}, x_i) plane. Indeed, the equation $f_i(x_{i-1}, x_i, x_{i+1}) = 0$ can be solved for

$$x_{i+1} = y_{i+1}(x_{i-1}, x_i),$$

since $\partial_{i+1} f_i > 0$. Similarly, since $\partial_{i-1} f_i > 0$, there is a function η_{i-1} with $x_{i-1} = \eta_{i-1}(x_i, x_{i+1})$ solving

$$f_i(\eta_{i-1}(x_i, x_{i+1}), x_i, x_{i-1}) = 0.$$

The assumption (H3) implies that both functions η_{i-1} and y_{i+1} are defined on the whole real line $\mathbf{R}$. We shall need this fact below in the construction of functions a_i. It is easy to see that $\partial_{i-1} y_{i+1}(x_i, x_{i-1}) < 0$ and

$$\partial_{i+1}\eta_{i-1}(x_i, x_{i+1}) < 0. \tag{4}$$

We now proceed with the induction step. We construct functions a_i, g_i and γ_i satisfying properties (A1)–(A2), (B) and (C) for $k := i$. The construction will be achieved in three steps. In the first step we will define a_i and verify properties (A1)–(A2). In the second step we will define g_i and show that (B) holds. The last step will be to check that $\gamma_i = \partial_{i+1}(a_i f_i)$ satisfies (C).

Step 1. *Construction of* a_i.

The function a_i must satisfy condition (A2) where the right-hand side $\partial_i(a_{i-1} f_{i-1})$ is the function γ_{i-1} already constructed in the previous step of the induction.

We first define a_i on the zero set $z(f_i)$ of f_i. On $z(f_i)$, which can be written as $x_{i+1} = y_{i+1}(x_{i-1}, x_i)$, the condition (A2) takes the form

$$\gamma_{i-1} = \partial_{i-1}(a_i f_i) = a_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)) \cdot \partial_{i-1} f_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)),$$

because $f_i = 0$. In order to satisfy (A2) we must therefore define the function a_i on the set $z(f_i)$ by the identity

$$a_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)) \cdot \partial_{i-1} f_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)) = \gamma_{i-1}(x_{i-1}, x_i). \tag{5}$$

To simplify notation we denote $\alpha_i(x_{i-1}, x_i) := a_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i))$. Observe that

$$\alpha_i > 0 \tag{6}$$

since $\partial_{i-1} f_i > 0$ by assumption (H1) and $\gamma_{i-1} > 0$ by induction hypothesis.

Now we want to define a_i outside the zero set $z(f_i)$ of f_i. We set

$$a_i(x_{i-1}, x_i, x_{i+1}) = \frac{1}{f_i} \int_{\eta_{i-1}(x_i, x_{i+1})}^{x_{i-1}} \partial_{i-1} f_i(\zeta_{i-1}, x_i, y_{i+1}(\zeta_{i-1}, x_i)) \alpha_i(\zeta_{i-1}, x_i) d\zeta_{i-1}. \tag{7}$$

Observe that this definition makes sense only for those (x_{i-1}, x_i, x_{i+1}) for which $y_{i+1}(\zeta_{i-1}, x_i)$ and $\eta_{i-1}(x_i, x_{i+1})$ are defined. By assumption (H3), these functions are defined for all $(x_{i-1}, x_i, x_{i+1}) \in \mathbf{R}^3$. So the definition of a_i does make sense, and a_i is defined for all $(x_{i-1}, x_i, x_{i+1}) \in \mathbf{R}^3$.

We now check properties (A1) and (A2) for a_i. We show first that a_i is continuous and that $\partial_{i-1}(a_i f_i)$ and $\partial_{i+1}(a_i f_i)$ exist and are continuous. These properties obviously hold at points (x_{i-1}, x_i, x_{i+1}) which do not belong to the zero set $z(f_i)$ of the function f_i. The calculation for the points in the set $z(f_i)$ is straightforward, but tedious. In order not to disrupt the argument we postpone the proof to the Appendix.

Now we show that a_i is positive. By assumption $\partial_{i-1} f_i > 0$ and by (6) also $\alpha_i > 0$. Furthermore, $f_i = f_i(x_{i-1}, x_i, x_{i+1})$ is positive for $x_{i-1} > \eta_i(x_i, x_{i+1})$, and negative when $x_{i-1} < \eta_{i-1}(x_i, x_{i+1})$. In the first case the right-hand side of (7) is positive since all the entries are positive; in the second case $f_i < 0$ but since the order of integration changes, the right-hand side is still positive. For $x_{i-1} = \eta_i(x_i, x_{i+1})$ the function a_i is positive by (6). Therefore $a_i > 0$ for all $(x_{i-1}, x_i, x_{i+1}) \in \mathbf{R}^3$. Thus a_i satisfies (A1).

Now we check condition (A2) at an arbitrary point (x_{i-1}, x_i, x_{i+1}). For any point $(x_{i-1}, x_i, x_{i+1}) \notin z(f_i)$ we have

$$\begin{aligned} \partial_{i-1}(a_i f_i) &\overset{(7)}{=} \partial_{i-1} f_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)) \alpha_i(x_{i-1}, x_i) \\ &= a_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)) \partial_{i-1} f_i(x_{i-1}, x_i, y_{i+1}(x_{i-1}, x_i)) \\ &\overset{(5)}{=} \gamma_{i-1}(x_{i-1}, x_i) \\ &= \partial_i(a_{i-1} f_{i-1}), \end{aligned} \tag{8}$$

where the last equality is the definition of γ_{i-1}. Since $\partial_{i-1}(a_i f_i)$ and $\partial_{i+1}(a_i f_i)$ are continuous, (8) holds for all (x_{i-1}, x_i, x_{i+1}). This verifies (A2).

Therefore a_i, as defined in (5, 7) indeed satisfies (A1)–(A2).

Step 2. *Construction of* g_i. The goal in this step is to define the function g_i such that (B) holds. Condition (B) requires that $\partial_i g_i = a_i f_i - \partial_i g_{i-1}$. The function g_{i-1} is already constructed, by the induction hypothesis, and $a_i f_i$ has been constructed in Step 1. Our expression for $\partial_i g_i$ does not depend on x_{i-1} since

$$\begin{aligned} \partial_{i-1}(a_i f_i - \partial_i g_{i-1}) &= \partial_i(a_{i-1} f_{i-1}) - \partial_{i-1}\partial_i g_{i-1} \\ &= \partial_i(\partial_{i-1} g_{i-1} + \partial_{i-1} g_{i-2}) - \partial_{i-1}\partial_i g_{i-1} \\ &= \partial_{i-1}\partial_i g_{i-2}(x_{i-2}, x_{i-1}) = 0. \end{aligned}$$

In the first line, we used that $\partial_{i-1}(a_i f_i) = \partial_i(a_{i-1} f_{i-1})$, which is (A2) for $k = i$ and was verified in (8). The induction hypothesis (B) for $k = i - 1$ gives that $a_{i-1} f_{i-1} =$

$\partial_{i-1}g_{i-1} + \partial_{i-1}g_{i-2}$. Also $\partial_{i-1}\partial_i g_{i-1} = \partial_i\partial_{i-1}g_{i-1}$ follows from the continuity of $\partial_i\partial_{i-1}g_{i-1}$, which is guaranteed by induction hypothesis (B) for $k = i-1$.

Therefore we define

$$g_i = g_i(x_i, x_{i+1}) := \int_0^{x_i} ((a_i f_i)(x_{i-1}, \zeta_i, x_{i+1}) - \partial_i g_{i-1}(x_{i-1}, \zeta_i))d\zeta_i.$$

Since $\partial_{i+1}\partial_i g_i = \partial_{i+1}(a_i f_i)$ and $\partial_{i+1}(a_i f_i)$ is continuous by step 1, the condition (B) holds for $k = i$.

Step 3. $\gamma_i > 0$.

The remaining step in the induction is to show (C) for $\gamma_i := \partial_{i+1}(a_i f_i)$, i.e., that $\gamma_i > 0$.

We differentiate

$$\begin{aligned}\gamma_i = \partial_{i+1}(a_i f_i) &= \partial_{i+1} \int_{\eta_{i-1}(x_i, x_{i+1})}^{x_{i-1}} \partial_{i-1} f_i(\zeta_{i-1}, x_i, y_{i+1}(\zeta_{i-1}, x_i))\alpha_i(\zeta_{i-1}, x_i)d\zeta_{i-1} \\ &= -\partial_{i+1}\eta_{i-1}(x_i, x_{i+1}) \cdot \partial_{i-1} f_i(\eta_{i-1}(x_i, x_{i+1}), x_i, y_{i+1}(\eta_{i-1}, x_i))\alpha_i(\eta_{i-1}(x_i, x_{i+1}), x_i) \\ &= -\partial_{i+1}\eta_{i-1}(x_i, x_{i+1}) \cdot \partial_{i-1} f_i(\eta_{i-1}(x_i, x_{i+1}), x_i, x_{i+1})\alpha_i(\eta_{i-1}(x_i, x_{i+1}), x_i)\end{aligned}$$

since $x_{i+1} = y_{i+1}(\eta_{i-1}(x_i, x_{i+1}), x_i)$. We see that the function γ_i is a function of x_i and x_{i+1} only. Furthermore, since $-\partial_{i+1}\eta_{i-1} > 0$ by (4), $\partial_{i-1}f_i > 0$ by assumption (H1), and $\alpha_i > 0$ by (6), we see that

$$\gamma_i = \partial_{i+1}(a_i f_i) > 0.$$

This finishes the induction step and thus proves existence of the Lyapunov function V. □

Proof of Theorem 1.2. The Lyapunov function V constructed in Theorem 1.1 is also a Lyapunov function for the system (2) on the closed positive orthant $\mathcal{O}$. Indeed,

$$\begin{aligned}\dot V &= -\sum_i \dot x_i(\partial_i g_i + \partial_i g_{i-1}) \\ &= -\sum_i \dot x_i a_i f_i \\ &= -\sum_i a_i x_i (f_i)^2 \le 0\end{aligned}$$

since $a_i > 0$ and $x_i \ge 0$ in the positive orthant $\mathcal{O}$. The derivative $\dot V = 0$ if and only if $x_i(f_i)^2 = 0$ for all i. This is equivalent to $x_i f_i = 0$ for all i. Since $\dot x_i = x_i f_i$, we have $\dot V = 0$ if and only if $\dot x_i = 0$ for all i. This implies that $\dot V(x) = 0$ if and only if x is an equilibrium. Therefore V is a Lyapunov function for (2). □

Proof of Theorem 1.3. Observe that the Lyapunov function $V = -\sum g_i(x_i, x_{i+1})$ is defined on the closed positive orthant $\mathcal{O}$. Since we assume (H2) for the functions f_i in (2), we see immediately that $|x_i| > C$ for the maximal $|x_i|$ implies $\dot x_i < 0$. It follows that each trajectory enters the positively invariant box $Q := \{x \in \mathbf{R}^n \mid |x_i| \le C\}$ in finite time. Since the positive orthant $\mathcal{O}$ is invariant under (2), each trajectory starting in $\mathcal{O}$ enters the set $\mathcal{O} \cap Q$ in finite time. By the LaSalle's invariance principle each trajectory in $\mathcal{O} \cap Q$ converges to the set where $\dot V = 0$, which is the set of equilibria.

So there is no chain recurrent set in $\mathcal{O}$, apart from the set of equilibria. In particular, there are no cycles of equilibria in $\mathcal{O}$. □

Remark 2.1. We assume (H3) in the definition of the function a_i; it guarantees that we can define a_i for all points (x_{i-1}, x_i, x_{i+1}), since for each such point the values $\eta_{i-i}(x_i, x_{i+1})$ and $y_{i+1}(x_{i-1}, x_i)$ are defined. Suppose now that (H3) is violated and that for some fixed $x_i = x_i^*$ the zero set $z(f_i, x_i^*)$ of f_i with $x_i = x_i^*$ fixed is a curve which satisfies

$$\lim_{x_{i-1}\to-\infty} z(f_i, x_i^*) = a_i^-(x_i^*) > -\infty, \qquad \lim_{x_{i-1}\to\infty} z(f_i, x_i^*) = a_i^+(x_i^*) < \infty.$$

Then the function a_i can only be defined in points (x_{i-1}, x_i, x_{i+1}) for which $a_i^-(x_i^*) < x_{i+1} < a_i^+(x_i^*)$. We also observe that the absolute value $|a_i|$ grows without bounds as (x_{i-1}, x_i, x_{i+1}) approaches the boundary of this region. Similar restrictions occur if we assume that

$$\lim_{x_{i+1}\to-\infty} z(f_i, x_i^*) = a_i^-(x_i^*) > -\infty, \qquad \lim_{x_{i+1}\to\infty} z(f_i, x_i^*) = a_i^+(x_i^*) < \infty.$$

However, observe that we do not need to define the Lyapunov function on all of $\mathbf{R}^n$. It follows from the dissipativeness assumption (H2) that each trajectory enters the box $Q := \{x \in \mathbf{R}^n \mid |x_i| \le C\}$ in finite time. We need to define our Lyapunov function V and thus functions a_i only for $x \in Q$. Thus we can still construct the Lyapunov function V provided that $|a_i^+(x_i^*)| > C$ and $|a_i^-(x_i^*)| > C$ for all i, with C given in assumption (H2).

Appendix. We show that the function a_i is continuous and the partial derivatives $\partial_{i-1}(a_i f_i)$ and $\partial_{i+1}(a_i f_i)$ exist and are continuous.

For all $(x_{i-1}, x_i, x_{i+1}) \notin z(f_i)$ this follows from the definition of a_i. We hence consider $(x_{i-1}, x_i, x_{i+1}) \in z(f_i)$.

We first prove continuity of a_i. Write $x_{i-1} = \eta_{i-1}(x_i, x_{i+1}) + h$ and expand f_i with respect to h at the point $(\eta_{i-1}(x_i, x_{i+1}), x_i, x_{i+1}) \in z(f_i)$:

$$\begin{aligned} &f_i(\eta_{i-1}(x_i, x_{i+1}) + h, x_i, x_{i+1}) \\ &\qquad = h(\partial_{i-1} f_i(\eta_{i-1}(x_i, x_{i+1}), x_i, x_{i+1}) + \tau(h, x_i, x_{i+1})). \end{aligned} \tag{9}$$

Here the remainder $\tau(h, x_i, x_{i+1})$ is a continuous function with $\tau(0, x_i, x_{i+1}) = 0$.

The integral in definition (7) of a_i becomes

$$\int_{\eta_{i-1}(x_i,x_{i+1})}^{\eta_{i-1}(x_i,x_{i+1})+h} \partial_{i-1} f_{i-1}(\zeta_{i-1}, x_i, y_{i+1}(\zeta_{i-1}, x_i))\alpha_i(\zeta_{i-1}, x_i) d\zeta_{i-1},$$

for $h \neq 0$ with a (uniformly) continuous integrand. Indeed, α_i, which is the restriction of a_i to $z(f_i)$, is continuous by definition (5). From the standard integration theory, we immediately obtain

$$\begin{aligned} \lim_{h\to0, h\neq0} \quad & \frac{1}{h}\int_{\eta_{i-1}(x_i,x_{i+1})}^{\eta_{i-1}(x_i,x_{i+1})+h} \partial_{i-1} f_{i-1}(\zeta_{i-1}, x_i, y_{i+1}(\zeta_{i-1}, x_i))\alpha_i(\zeta_{i-1}, x_i) d\zeta_{i-1} \\ &= (a_i \partial_{i-1} f_i)(\eta_{i-1}(x_i, x_{i+1}), x_i, y_{i+1}(\eta_{i-1}, x_i)) \\ &= (a_i \partial_{i-1} f_i)(\eta_{i-1}(x_i, x_{i+1}), x_i, x_{i+1}). \end{aligned}$$

The limit is locally uniform with respect to x_i, x_{i+1} by continuity of all functions involved. Inserting the expansion (9) of f_i, we obtain

$$\lim_{h\to 0,h\neq 0} \quad \frac{1}{f_i}\int_{\eta_{i-1}(x_i,x_{i+1})}^{\eta_{i-1}(x_i,x_{i+1})+h} \partial_{i-1}f_{i-1}(\zeta_{i-1},x_i,x_{i+1})\alpha_i(\zeta_{i-1},x_i)d\zeta_{i-1}$$

$$= \lim_{h\to 0,h\neq 0} \frac{h\cdot(a_i\partial_{i-1}f_i)(\eta_{i-1}(x_i,x_{i+1}),x_i,x_{i+1})}{h\cdot(\partial_{i-1}f_i(\eta_{i-1}(x_i,x_{i+1}),x_i,x_{i+1})+\tau(h,x_i,x_{i+1}))}$$

$$= a_i(\eta_{i-1}(x_i,x_{i+1}),x_i,x_{i+1}),$$

locally uniformly with respect to x_i, x_{i+1}. This proves the continuity of a_i, defined by (5) and (7).

Now we show that the partial derivatives $\partial_{i-1}(a_if_i)$ and $\partial_{i+1}(a_if_i)$ exist and are continuous.

This is immediate from the integral representation (7) of a_if_i, which holds for all (x_{i-1},x_i,x_{i+1}); from continuity of the integrand, which does not contain x_{i+1} and x_{i-1}; from differentiability of η_{i-1}; and from the fundamental theorem of calculus.

Acknowledgments. T.G. wants to thank Mary Silber and Vivien Kirk for the useful discussion, which showed him the futility of the previous attempt to solve the problem.

REFERENCES

[FG] B. Fiedler and T. Gedeon, *A class of convergent neural network dynamics*, Phys. D, 111 (1998), pp. 288–294.

[HK] J. Hale and H. Koçak, *Dynamics and Bifurcations*, Springer-Verlag, Berlin, 1991.

[Hi1] M. Hirsch, *Systems of differential equations which are competitive or cooperative* I: *Limit sets*, SIAM J. Math. Anal., 13 (1982), pp. 167–179.

[Hi2] M. Hirsch, *Systems of differential equations that are competitive or cooperative* II: *Convergence almost everywhere*, SIAM J. Math. Anal., 16 (1985), pp. 432–439.

[Hi3] M. Hirsch, *Systems of differential equations which are competitive or cooperative* III: *Competing species*, Nonlinearity, 1 (1988), pp. 51–71.

[Hi4] M. Hirsch, *Stability and convergence in strongly monotone dynamical systems*, J. Reine Angew. Math., 383 (1988), pp. 1–53.

[F-S] H. I. Freedman and H. L. Smith, *Tridiagonal competitive-cooperative Kolmogorov systems*, Differential Equations Dynam. Systems, 3 (1995), pp. 367–382.

[M] H. Matano, *Strong comparison principle in nonlinear parabolic equations*, in Nonlinear Parabolic Equations: Qualitative Properties of Solutions, L. Boccardo and A. Tesei, eds., Pitman Res. Notes in Math., Ser. 149, Longman Scientific and Technical, London, 1987, pp. 148–155.

[P] P. Poláčik, *Convergence in smooth strongly monotone flows defined by semilinear parabolic equations*, J. Differential Equations, 79 (1989), pp. 89–110.

[Sm] J. Smillie, *Competitive and cooperative tridiagonal systems of differential equations*, SIAM J. Math. Anal, 15 (1984), pp. 530–534.

[S1] H. L. Smith, *Periodic tridiagonal competitive and cooperative systems of differential equations*, SIAM J. Math. Anal., 22 (1991), pp. 1102–1109.

[S2] H. L. Smith, *Monotone Dynamical Systems*, Math. Surveys Monogr. 41, AMS, Providence, RI, 1995.

[ST1] H. L. Smith and H. Thieme, *Quasi convergence and stability for strongly order-preserving semiflows*, SIAM J. Math. Anal., 21 (1990), pp. 673–692.

[ST2] H. L. Smith and H. Thieme, *Convergence for strongly order-preserving semiflows*, SIAM J. Math. Anal, 22 (1991), pp. 1081–1100.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 479–496

LOGARITHMIC CONVEXITY AND THE "SLOW EVOLUTION" CONSTRAINT IN ILL-POSED INITIAL VALUE PROBLEMS*

ALFRED S. CARASSO†

Abstract. This paper examines a wide class of ill-posed initial value problems for partial differential equations, and surveys logarithmic convexity results leading to Hölder-continuous dependence on data for solutions satisfying prescribed bounds. The discussion includes analytic continuation in the unit disc, time-reversed parabolic equations in L^p spaces, the time-reversed Navier–Stokes equations, as well as a large class of nonlocal evolution equations that can be obtained by randomizing the time variable in abstract Cauchy problems. It is shown that in many cases, the resulting Hölder-continuity is too weak to permit useful continuation from imperfect data. However, considerable reduction in the growth of errors occurs, and continuation becomes feasible, for solutions satisfying the *slow evolution from the continuation boundary* constraint, previously introduced by the author.

Key words. ill-posed problems, analytic continuation, backwards in time continuation, logarithmic convexity, Hölder-continuity, parabolic equations, growing diffusion coefficients, non self-adjoint problems, Navier–Stokes equations, holomorphic semigroups, subordinated processes, slow evolution from continuation boundary, SECB constraint

AMS subject classifications. 35R25, 35B60, 35B35, 47D06

PII. S0036141098332366

1. Introduction.

The problem of reconstructing the past behavior of a system, given knowledge of its current state, is of interest to many branches of science. For evolution equations, such backwards in time continuation is typically ill-posed in the presence of dissipative terms. Other spatial continuation problems in elliptic or parabolic equations exhibit similar characteristics. This paper is concerned with the Hölder-continuous dependence on data that results when certain ill-posed continuation problems in partial differential equations are stabilized by prescribed bounds [14], [21].

Because the Hölder exponent must decay to zero as the continuation boundary is approached, there is an unavoidable growth of errors originating from imperfect data. In some cases, such errors may preclude continuation into a region of particular interest. The *slow evolution from the continuation boundary* (SECB) constraint introduced in [6], [7] is an a priori statement about the rate of change of the desired solution near the continuation boundary. This information supplements information provided by prescribed bounds on the solution. As a consequence, stronger stability estimates can be obtained for solutions satisfying the SECB constraint than is otherwise possible. This constraint was shown to be effective in controlling the growth of noise in certain *image deblurring* problems, in which backwards in time continuation in diffusion equations involving fractional Laplacians plays a key role. In these problems, the Hölder exponent decays *linearly* to zero.

The present self-contained paper deals with a much wider class of problems. We survey important classes of equations, including the Navier–Stokes equations, where logarithmic convexity inequalities can be shown to hold. Using the theory of holomorphic semigroups, we consider parabolic equations in L^p spaces, as well as a large class of nonparabolic problems, typically involving nonlocal differential operators, that can

*Received by the editors January 9, 1998; accepted for publication (in revised form) August 6, 1998; published electronically March 19, 1999.
http://www.siam.org/journals/sima/30-3/33236.html
†Mathematical and Computational Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899 (alfred@cam.nist.gov).

be constructed by *subordination* [8]. The resulting Hölder exponents are particularly instructive in their dependence on the continuation variable. Linear decay to zero is the exception, the behavior being generally sublinear, and rapid exponential decay is possible in some cases. In time-dependent problems, such Hölder exponents are indicative of the rate at which the evolution equation has forgotten the past, and hence, of the subsequent difficulty of reconstructing the past from imperfect knowledge of the present.

It is shown that the SECB constraint, when applicable, becomes progressively more significant the faster the Hölder exponent decays to zero. Considerable error reduction is possible in many cases. Indeed, continuation problems that were heretofore intractable may become amenable to numerical computation, provided their solutions satisfy an SECB constraint. The paper concludes with a simple, explicit example of backwards in time continuation in an evolution equation with exponentially decaying Hölder exponent.

The following problem is important in its own right and serves to motivate the subsequent discussion.

1.1. Analytic continuation in the unit disc. Let $\mathcal{A}$ be the class of complex-valued functions $u(z)$ that are continuous in the closed unit disc and holomorphic in its interior, and let

$$\| u(r) \|_\infty = \max_{0\le\theta\le 2\pi} |u(re^{i\theta})|. \tag{1}$$

Fix a with $0 < a < 1$, and consider the problem of determining $u(re^{i\theta})$ for $a < r < 1$ from approximate knowledge of $u(z)$ on the circle $r = a$. Hadamard's three-circle theorem asserts that $\log \| u(r) \|_\infty$ is a convex function of $\log r$ for $a \le r \le 1$. If

$$\omega(r) = \log r / \log a, \qquad 0 < a \le r \le 1, \tag{2}$$

then

$$\| u(r) \|_\infty \le \| u(1) \|_\infty^{1-\omega(r)} \| u(a) \|_\infty^{\omega(r)}, \qquad 0 < a \le r \le 1. \tag{3}$$

We have equality in (3) for $u(z) = z^n$. This convexity inequality is the basis for stabilizing the ill-posed continuation problem when noisy data are given on $r = a$. Restrict the class of admissible continuations to functions $u(z) \in \mathcal{A}$ satisfying a *prescribed* bound,

$$\| u(1) \|_\infty \le M. \tag{4}$$

Fix $\epsilon > 0$, $\epsilon \ll M$, and let data $f(\theta)$ be given on $r = a$ such that for some $u(z) \in \mathcal{A}$ satisfying (4), we have

$$\| u(a) - f \|_\infty \le \epsilon. \tag{5}$$

If now $u_1(z), u_2(z) \in \mathcal{A}$ are any two objects satisfying (4) and (5), we get from (3)

$$\| u_1(r) - u_2(r) \|_\infty \le 2M^{1-\omega(r)} \epsilon^{\omega(r)}, \qquad a \le r \le 1. \tag{6}$$

For fixed $r_0 < 1$, the difference between any two possible continuations at $r = r_0$ can be made arbitrarily small in the L^∞ norm by giving sufficiently accurate data at $r = a$, i.e., by making $\epsilon > 0$ *sufficiently small* in (5). On the other hand, no

matter how small one chooses ϵ in (5), the inequality (6) cannot ensure accurate results at the continuation boundary $r = 1$, since $\omega(1) = 0$. Indeed, with M given in (4), and $\epsilon > 0$ given in (5), consider continuing the function $u(z) \equiv M/2$ from data $f(\theta) = M(1 + a^n e^{in\theta})/2$ at $r = a$, where n is such that $a^n < 2\epsilon/M$. At $r = 1$, the continuation $v(z) = M(1+z^n)/2$ satisfies the prescribed bound, but approximates the desired continuation $u(z) \equiv M/2$ with a relative error of 100% in the L^∞ norm. The inequality (6) establishes Hölder-continuous dependence on the data only on compact subsets of the region where bounds are prescribed. This situation prevails in diverse classes of improperly posed problems in partial differential equations stabilized by a priori bounds. Use of such bounds, together with the analysis of the resulting continuity with respect to the data, was pioneered by Fritz John in a landmark paper [14].

A basic difficulty with the above Hölder-continuity is the following. In most applications, $\epsilon > 0$ is determined by the accuracy of the instrumentation used to acquire the data. While ϵ is usually small, it is *fixed* and *cannot* be made arbitrarily small. In such applications, the dependence of the Hölder exponent $\mu(t)$ on the continuation variable t plays a crucial role. In some cases, such as backwards in time continuation in the heat equation, we have $\mu(t) = t/T$, so that $\mu(t)$ decays linearly to zero as continuation progresses from $t = T > 0$ to the continuation boundary $t = 0$. More typically, μ is *sublinear* in the continuation variable. If μ decays too rapidly to zero, useful continuation becomes impossible, even in regions well away from the continuation boundary. This is the case in (2), for example, when $a > 0$ is small. In the case of evolution equations, as will be seen below, rapid decay of μ to zero can be brought about by various factors, including nonlinearity, non-self-adjointness, diffusion coefficients that grow with time, or adverse spectral properties in the spatial differential operator.

It develops that while prescribed bounds are necessary to stabilize ill-posed initial value problems, they are frequently insufficient to allow continuation far enough into the region of interest. Further a priori information must be provided for this purpose. In this paper we show that knowing the rate of change of the desired solution near the continuation boundary is *slow* can be very helpful.

2. Slow evolution from the continuation boundary (SECB). We consider linear or nonlinear continuation problems in a single variable t, $0 \le t \le T$, with continuation boundary at $t = 0$. In spatial continuation problems with radial symmetry, t is a radial coordinate, e.g., $t = 1 - r$ in (1). In applications involving continuation in the time variable, t is related to time. In an appropriate Banach space X with norm $\| \quad \|$, the continuation $u(t)$ is an X-valued function with norm $\| u(t) \|$ for fixed t. Let $u_1(t)$, $u_2(t)$, be any two continuations from the given data $f(x)$ at $t = T$, with $\| u_i(T) - f \| \le \epsilon$, and satisfying a prescribed bound, $\| u_i(0) \| \le M$ at $t = 0$. Here, ϵ, $M > 0$ are both known, and $\epsilon \ll M$. Let $w(t) = u_1(t) - u_2(t)$. We assume $w(t)$ satisfies a convexity inequality

$$\| w(t) \| \le \| w(0) \|^{1-\mu(t)} \| w(T) \|^{\mu(t)}, \qquad 0 \le t \le T \tag{7}$$

with known exponent $\mu(t)$, $0 \le \mu(t) \le 1$. For given K with $0 < K \ll M/\epsilon$, define μ^* by

$$\mu^* = \log\{M/(M - K\epsilon)\}/\log(M/\epsilon). \tag{8}$$

The SECB constraint is expressed as follows: *There exists a known small constant $K > 0$, and a known fixed $s > 0$, with $\mu(s) > \mu^*$, such that* $\| u(s) - u(0) \| \le K\epsilon$.

By continuity as $t \downarrow 0$, given any $\epsilon > 0$, there always exists a sufficiently small $s > 0$ such that the last inequality holds with a small K. However, the requirement that s be known and be such that $\mu(s) > \mu^*$, constitutes further a priori information about the continuation problem. It will turn out to be desirable that $\mu(s) \gg \mu^*$.

There are several sets of circumstances that can result in solutions satisfying SECB. As an example, consider linear parabolic initial value problems with time-independent coefficients, homogeneous boundary conditions, and no forcing term. If the coefficients are small and the initial values are not dominated by very high frequency components, the solution will evolve slowly from these data at the continuation boundary $t = 0$. This situation prevails in some important biomedical image deblurring problems, where the blurring kernel is a Gaussian distribution with small variance. In that case, the blurred image can be viewed as the solution at time $t = T > 0$, of an initial value problem for a diffusion equation with a small diffusivity, the data at $t = 0$ being the desired unblurred image. See [6], [7, Fig. 2]. Despite the small diffusivity, fine scale information that may be of vital significance typically cannot be discerned in the blurred image. Hence the need for deblurring. More generally, in parabolic problems with time-dependent coefficients, consider the case where the coefficients are initially small but grow with time. Again, the solution will evolve slowly from the initial values, while it may change rapidly at later times. See the example in section 8 below. Inhomogeneous boundary conditions provide another mechanism that can produce solutions satisfying SECB, even when the coefficients are not small. Thus, if a body in thermal equilibrium at $t = 0$ is subjected to a boundary heat flux $b(t)$, where, with $b(0) = 0$, $b(t)$ increases slowly in the interval $0 \le t \le T/4$; increases rapidly between $T/4$ and $T/2$; and decreases rapidly to zero between $T/2$ and $3T/4$, the solution at time T will differ considerably from its initial values, while evolving slowly near $t = 0$. Similar behavior can occur in the Navier–Stokes initial value problem. Consider flows in lid-driven cavities as in [20] and the references therein. If the velocity of the driving lid has a time dependence similar to that in the heat flux $b(t)$ above, the solution of the Navier–Stokes system at time $T > 0$ will differ substantially from its initial state, while having evolved slowly near $t = 0$. Examples of SECB may likewise be found in spatial continuation problems.[1]

LEMMA 1. *For $1 > \mu(s) > \mu^*$, let $\Gamma(K, s)$ be the unique root of the transcendental equation*

$$x = K + x^{1-\mu(s)}. \tag{9}$$

Then

$$\begin{aligned} &K + 1 < \Gamma < M/\epsilon, \\ &\{K/\mu(s)\} \le \Gamma \log \Gamma \le \{K/\mu(s)\}\{\Gamma/(\Gamma - K)\}, \\ &\Gamma \log \Gamma \approx K/\mu(s) \le \{\mu^*/\mu(s)\}(M/\epsilon) \log M/\epsilon, \qquad K \ll \Gamma. \end{aligned} \tag{10}$$

Moreover, if $K + 1 \le x_0 \le M/\epsilon$, the iteration $x_{n+1} = K + x_n^{1-\mu(s)}$ converges to Γ.

Proof. The curve $y = x$ intersects the curve $y = K + x^{1-\mu(s)}$ at a single point, Γ. From (8), we have $M/\epsilon = K + (M/\epsilon)^{1-\mu^*}$, so that M/ϵ is the root of (9) when

[1] In spatial continuation for the heat equation in the quarter plane, or *sideways heat equation* problem [10], large, rather than small, diffusivities near $x = 0$ are conducive to slow evolution from that boundary.

$\mu(s) = \mu^*$. The roots of (9) decrease monotonically as $\mu(s)$ increases. Therefore, $\Gamma < M/\epsilon$. Evidently, $\Gamma > 1$, which implies $\Gamma > K + 1$. Using the inequality $w \le \log\{1/(1-w)\} \le w/(1-w)$, $0 \le w < 1$, we get $K/\Gamma \le \mu(s)\log\Gamma \le K/(\Gamma - K)$. Thus, $\Gamma\log\Gamma \approx K/\mu(s)$ if $K \ll \Gamma$. Next, $K\epsilon/M \le \mu^*\log(M/\epsilon)$, which leads to the last inequality in (10). The last statement in Lemma 1 is a standard result called "fixed point iteration." □

THEOREM 1. *Let ϵ, M, K be given positive constants with $\epsilon < M$ and $K\epsilon < M$. Let X be a Banach space with norm $\| \ \|$, and let $f \in X$. Let $\mathcal{C}$ be a linear or nonlinear continuation problem from the data f for the X-valued function $u(t)$, $0 \le t \le T$, where $\| u(0) \| \le M$ and $\| u(T) - f \| \le \epsilon$. Let $\mathcal{C}$ be such that the difference $w(t)$ of any two possible continuations satisfies*

$$\| w(t) \| \le \| w(0) \|^{1-\mu(t)} \| w(T) \|^{\mu(t)}, \qquad 0 \le t \le T, \tag{11}$$

with known $\mu(t)$, $0 \le \mu(t) \le 1$. If the solutions of $\mathcal{C}$ also satisfy $\| u(s) - u(0) \| \le K\epsilon$ for some known $s > 0$ with $\mu(s) > \mu^*$, where μ^* is defined in (8), then

$$\| w(t) \| \le 2\Gamma^{1-\mu(t)}\epsilon, \qquad 0 \le t \le T, \tag{12}$$

where Γ is the constant defined in Lemma 1. Moreover, $\Gamma \ll M/\epsilon$ if $\mu^* \ll \mu(s)$.

Proof. From (11), the difference of any two continuations satisfies

$$\begin{aligned} &\| w(t) \| \le \Lambda^{1-\mu(t)}\delta^{\mu(t)}, \qquad 0 \le t \le T, \\ &\| w(s) - w(0) \| \le K\delta, \qquad s > 0, \quad \mu(s) > \mu^*, \end{aligned} \tag{13}$$

where $\Lambda = 2M$, $\delta = 2\epsilon$. From

$$\begin{aligned} \| w(t) \| &\le \| w(0) \|^{1-\mu(t)} \| w(T) \|^{\mu(t)} \\ &\le \{\| w(s) - w(0) \| + \| w(s) \|\}^{1-\mu(t)} \| w(T) \|^{\mu(t)}, \end{aligned} \tag{14}$$

together with (13), we get

$$\| w(s) \| \le \left\{K\delta + \Lambda^{1-\mu(s)}\delta^{\mu(s)}\right\}^{1-\mu(s)}\delta^{\mu(s)}. \tag{15}$$

The initial estimate for $w(s)$, $\| w^1(s) \| \le \Lambda^{1-\mu(s)}\delta^{\mu(s)}$, has been used in (14) to produce a new estimate, $\| w^2(s) \|$, given by (15). We may insert (15) back into (14) to produce a third estimate for $w(s)$, and so on. At the nth step of that iteration, we get $\| w^n(s) \| \le Z_n^{1-\mu(s)}\delta^{\mu(s)}$, where

$$Z_1 = \Lambda, \qquad Z_k/\delta = K + (Z_{k-1}/\delta)^{1-\mu(s)}, \qquad k > 1. \tag{16}$$

We have $Z_n \to \Gamma\delta$ as $n \uparrow \infty$, where Γ is defined in Lemma 1. Thus, $\| w(s) \| \le \Gamma^{1-\mu(s)}\delta$. Inserting this back into (14), and using (13), we get

$$\| w(t) \| \le 2\Gamma^{1-\mu(t)}\epsilon, \qquad 0 \le t \le T. \tag{17}$$

Finally, the last inequality in (10) shows that $\Gamma\log\Gamma \ll (M/\epsilon)\log(M/\epsilon)$ provided $\mu(s) \gg \mu^*$. □

To illustrate the SECB constraint, we return to analytic continuation in the unit disc, as discussed in the Introduction. Let $M = 4$, $\epsilon = 10^{-5}$, $a = 0.1$, and consider

continuing the function $u(z) = 1 + 0.1z$ from data $f(\theta) = 1 + 0.1ae^{i\theta} + 2a^6e^{6i\theta}$ at $r = a$. Let $v(z) = 1 + 0.1z + 2z^6$. Then,

$$\begin{aligned} &\| u(1) \|_\infty = 1.1 < M, && \| u(a) - f \|_\infty = 2 \times 10^{-6} < \epsilon, \\ &\| v(1) \|_\infty = 3.1 < M, && \| v(a) - f \|_\infty = 0 < \epsilon. \end{aligned} \tag{18}$$

Thus, both $u(z)$ and $v(z)$ satisfy the a priori constraints (4), (5). However, at $r = 3/4$, we find $\| u(3/4) - v(3/4) \|_\infty / \| u(3/4) \|_\infty = 33\%$, and $\| u(1) - v(1) \|_\infty / \| u(1) \|_\infty = 182\%$ at $r = 1$. These are unacceptable relative errors. Additional a priori information about $u(z)$ near the continuation boundary $r = 1$ can reduce this uncertainty. With $K = 12$ and $s = 0.001$, we have

$$\| u(1) - u(1 - s) \|_\infty = 0.1s \le K\epsilon, \tag{19}$$

while

$$\begin{aligned} \| v(1) - v(1 - s) \|_\infty &= \max_\theta |0.1se^{i\theta} + 2\{1 - (1 - s)^6\}e^{6i\theta}| \\ &= \max_\theta |10^{-4}e^{i\theta} + 1.197 \times 10^{-2}e^{6i\theta}| \\ &> 10^{-2} > K\epsilon. \end{aligned} \tag{20}$$

Therefore, the SECB constraint (19), with $s = 0.001$ and $K = 12$, eliminates $v(z)$ as a possible continuation. With $\omega(r)$ as in (2), let $\mu(t) = \omega(1 - t)$, $0 \le t \le 1 - a$. Since $\mu(s) = 4.345 \times 10^{-4}$, while $\mu^* = 2.326 \times 10^{-6}$, we have $\mu(s)/\mu^* = 187$, and $\Gamma \log \Gamma \approx K/\mu(s) = 27618$. This gives $\Gamma = 3397$ while $M/\epsilon = 400,000 = 118\Gamma$. Let $w(r, \theta)$ be the difference between any two possible continuations from data at $r = a = 0.1$ satisfying (4), (5), with $M = 4$ and $\epsilon = 10^{-5}$. Without the SECB constraint (19), we have $\| w(1) \|_\infty \le 2M = 8$. With the SECB constraint, we have $\| w(1) \|_\infty \le 2\ \Gamma\ \epsilon = 0.0679$.

Theorem 1 leads to the following corollary to the Hadamard three-circle theorem.

THEOREM 2 (corollary). *In the analytic continuation problem in the unit disc, let $u_1(z)$, $u_2(z)$ be as in (6), let $0 < s < 1 - a$, let $\omega(r)$ be as in (2), and let $\mu(s) = \omega(1 - s)$. If*

$$\| u_i(1) - u_i(1 - s) \|_\infty \le K\epsilon, \qquad i = 1, 2, \tag{21}$$

with known K, $0 < K < M/\epsilon$, and known s such that $\mu(s) > \mu^*$, where μ^* is defined in (8), then

$$\| u_1(r) - u_2(r) \|_\infty \le 2\Gamma^{1-\omega(r)}\epsilon, \qquad a \le r \le 1, \tag{22}$$

where $\Gamma < M/\epsilon$ is the constant in Lemma 1. Moreover, $\Gamma \ll M/\epsilon$ if $\omega(1 - s) \gg \mu^*$.

Remark 1. The SECB constraint does not imply differentiability of $u(1, \theta)$, as a function of θ, on the circle $r = 1$. More generally, at the continuation boundary $t = 0$, $u(t)$ need not be differentiable in its remaining variables in order to satisfy an SECB constraint. This point is emphasized in [6], [7, Fig. 2], and again in the example in section 8 below.

3. An approach to logarithmic convexity. The following method has been widely used to obtain continuous dependence inequalities in linear and nonlinear initial value problems, typically in a Hilbert space setting [2], [16], [17]. Let H be a

Hilbert space, and let $\mathcal{S}$ be an initial value problem for a system of partial differential equations, with solutions $u(t) \in H$ for each $t \in (0, T]$. Let $F(t)$ be a real-valued twice continuously differentiable function of t, defined on the set of solutions $u(t)$ of $\mathcal{S}$ and satisfying

$$\begin{aligned} &F(t) \geq 0, \qquad F(t) = 0 \Longleftrightarrow u(t) = 0, \qquad 0 \leq t \leq T, \\ &F(t)F''(t) - \{F'(t)\}^2 \geq -a_1\ F(t)F'(t) - a_2\ F^2(t), \qquad 0 < t < T, \end{aligned} \tag{23}$$

where a_1 and a_2 are constants. If $a_1 = a_2 = 0$ in (23), then

$$F(t) \leq \{F(0)\}^{(T-t)/T}\{F(T)\}^{t/T}, \qquad 0 \leq t \leq T. \tag{24}$$

More typically, $a_1 \neq 0$ in (23). In that case, let

$$m = -a_2/a_1, \qquad \mu(t) = \{e^{-a_1 t} - 1\}\{e^{-a_1 T} - 1\}^{-1}, \qquad 0 \leq t \leq T. \tag{25}$$

Then (see [2], [17])

$$e^{-mt}F(t) \leq \{F(0)\}^{1-\mu(t)}\{e^{-mT}F(T)\}^{\mu(t)}, \qquad 0 \leq t \leq T. \tag{26}$$

We now give two examples of the use of this technique in L^2, with $F(t) = \| u(t) \|^2$. Many other examples, and choices for $F(t)$, may be found in [2], [16], [24], and the references therein.

4. Self-adjoint parabolic problems with time-dependent coefficients. Let Ω be a bounded domain in R^n with sufficiently smooth boundary $\partial\Omega$. For $x \in R^n$ and $t \geq 0$, let $a(t; u, v)$ and $\dot{a}(t; u, v)$ be *symmetric* bilinear forms on $H_0^m(\Omega)$ given by

$$\begin{aligned} a(t; u, v) &= \sum_{|p|,|q| \leq m} \int_\Omega a_{pq}(x, t) D^q u \overline{D^p v} dx, \\ \dot{a}(t; u, v) &= \sum_{|p|,|q| \leq m} \int_\Omega \dot{a}_{pq}(x, t) D^q u \overline{D^p v} dx, \end{aligned} \tag{27}$$

where the coefficients a_{pq} depend smoothly on x and t, $\overline{a}_{pq} = a_{qp}$, and $\dot{a}_{pq}$ denotes $\partial a_{pq}/\partial t$. We assume $a(t; u, v)$ to be uniformly strongly coercive on $H_0^m(\Omega)$, i.e., there exists $\alpha > 0$, independent of t, such that

$$a(t; v, v) \geq \alpha \parallel v \parallel_m^2, \qquad v \in H_0^m(\Omega). \tag{28}$$

Both $a(t; u, v)$ and $\dot{a}(t, u, v)$ are continuous on $H_0^m(\Omega) \times H_0^m(\Omega)$, uniformly in t; i.e., there exist $\beta,\ \gamma > 0$, independent of t, such that

$$|a(t; u, v)| \leq \beta \parallel u \parallel_m \parallel v \parallel_m, \quad |\dot{a}(t; u, v)| \leq \gamma \parallel u \parallel_m \parallel v \parallel_m, \quad u, v \in H_0^m(\Omega). \tag{29}$$

The bilinear form $a(t, u, v)$ defines a positive self-adjoint operator $A(t)$ in $L^2(\Omega)$, [28], with domain $D_A = H^{2m}(\Omega) \cap H_0^m(\Omega)$ such that

$$\begin{aligned} &(A(t)v, v) = a(t : v, v), \qquad (\dot{A}(t)v, v) = \dot{a}(t; v, v), \qquad v \in D_A, \\ &|(\dot{A}(t)v, v)| \leq (\gamma/\alpha)(A(t)v, v), \qquad v \in D_A, \end{aligned} \tag{30}$$

where (,) denotes the scalar product in $L^2(\Omega)$. The operator $A(t)$ is the closed extension of the strongly elliptic symmetric differential operator

$$A(x,t,D)u = \sum_{|p|,|q|\le m} (-1)^{|p|} D^p(a_{pq}(x,t)D^q u), \qquad x \in \Omega, \quad t > 0, \tag{31}$$

with zero Dirichlet data on $\partial\Omega$. We distinguish two cases: a) the case where diffusion is constant or *decreases* with time and $(\dot{A}(t)v, v) \le 0$, and b) the case where diffusion *increases* at least some of the time and $(\dot{A}(t)v, v) \le \gamma \parallel v \parallel_m^2$. The theorem below is due to Agmon and Nirenberg [1].

THEOREM 3. *Let α and γ be the positive constants in* (28) *and* (29). *Let $u(t) \in L^2(\Omega)$ be a solution of $u_t = -A(t)u$, $t > 0$. If $(\dot{A}(t)v, v) \le 0$, $0 < t \le T$, let $\mu(t) = t/T$. If $(\dot{A}(t)v, v) \le \gamma \parallel v \parallel_m^2$, $0 < t \le T$, let $c = \gamma/\alpha$ and let $\mu(t) = \{e^{ct} - 1\}\{e^{cT} - 1\}^{-1}$. Then,*

$$\parallel u(t) \parallel \le \parallel u(0) \parallel^{1-\mu(t)} \parallel u(T) \parallel^{\mu(t)}, \qquad 0 \le t \le T. \tag{32}$$

Proof. With $F(t) = \parallel u(t) \parallel^2$, we have $F'(t) = -2(A(t)u, u)$, and

$$\begin{aligned} FF'' - \{F'\}^2 &= -2(\dot{A}(t)u, u)F + 4 \parallel A(t)u \parallel^2 \parallel u \parallel^2 - 4|(A(t)u, u)|^2 \\ &\ge -2(\dot{A}(t)u, u)F \end{aligned} \tag{33}$$

on using Schwarz's inequality. If $(\dot{A}(t)u, u) \le 0$, we have the case $a_1 = a_2 = 0$ in (23) and the result follows from (24). If $(\dot{A}(t)u, u) \le \gamma \parallel u \parallel_m^2$, we use (30) to obtain $-2(\dot{A}(t)u, u)F \ge (\gamma/\alpha)FF'$ in (33). This is the case $a_2 = 0$, $a_1 = -\gamma/\alpha$ in (23), and the result follows from (26) with $m = 0$. □

Evidently, growing diffusion coefficients can result in exponential decay in $\mu(t)$. In section 8 below, we study a simple explicit example where this is indeed the case. Applying Theorem 1 to the Agmon–Nirenberg result, we have the following corollary.

THEOREM 4 (corollary). *In Theorem* 3, *let positive constants ϵ, M, be given, with $\epsilon < M$. Let $f \in L^2(\Omega)$ be given data at time $T > 0$, and let $u_1(t)$, $u_2(t)$ be two solutions of $u_t = -A(t)u + g(t)$, $0 < t \le T$, such that $\parallel u_i(T) - f \parallel \le \epsilon$, and $\parallel u_i(0) \parallel \le M$, $i = 1, 2$. Let $w(t) = u_1(t) - u_2(t)$. Then, with $\mu(t)$ as in Theorem* 3,

$$\parallel w(t) \parallel \le 2M^{1-\mu(t)}\epsilon^{\mu(t)}, \qquad 0 \le t \le T. \tag{34}$$

If, in addition, $\parallel u_i(s) - u_i(0) \parallel \le K\epsilon$, $i = 1, 2$, with known K, $0 < K < M/\epsilon$, and known $s > 0$ such that $\mu(s) > \mu^*$, where μ^* is defined in (8), *then*

$$\parallel w(t) \parallel \le 2\Gamma^{1-\mu(t)}\epsilon, \qquad 0 \le t \le T, \tag{35}$$

where $\Gamma < M/\epsilon$ is the constant in Lemma 1. *Moreover, $\Gamma \ll M/\epsilon$ if $\mu(s) \gg \mu^*$.*

5. Navier–Stokes equations backwards in time. With $i = 1, 3$, and summation convention understood, consider the Navier–Stokes system in a bounded domain $\Omega \subset R^3$, with smooth boundary $\partial\Omega$,

$$\left.\begin{aligned} u_{i,t} &= \nu\Delta u_i - u_j u_{i,j} - \rho^{-1}p_{,i} + G_i(x,t) \\ u_{j,j} &= 0 \end{aligned}\right\} \qquad (x,t) \in \Omega \times (0,T], \tag{36}$$

$$u_i(x,T) = f_i(x), \quad x \in \Omega, \qquad u_i = g_i(x,t), \quad (x,t) \in \partial\Omega \times [0,T].$$

Here, differentiation is denoted by a comma, ν is the kinematic viscosity, ρ is the constant density, p the unknown pressure, $u_i(x,t)$ is the ith component of fluid velocity,

$G_i(x,t)$ is a prescribed body force per unit mass, and $g_i(x,t)$ are prescribed boundary values. In [18], Knops and Payne study the stability of reconstructing the solution of (36) on $[0,T)$, under small perturbations of the solution values $f_i(x)$ at some positive time T. Let P and Q be prescribed positive constants. A function $u_i(x,t)$ is said to belong to the set $\mathcal{P}$ provided

$$\sup_{\Omega\times[0,T]} u_i u_i \le P^2, \tag{37}$$

while it belongs to the set $\mathcal{Q}$ whenever

$$\sup_{\Omega\times[0,T]} \{u_i u_i + (u_{i,j}-u_{j,i})(u_{i,j}-u_{j,i}) + u_{i,t}u_{i,t}\} \le Q^2. \tag{38}$$

In [25], the same stability problem is studied under weaker constraints. Let $u_i^1(x,t)$ and $u_i^2(x,t)$ denote classical solutions of (36) corresponding to terminal data $f_i^1(x)$ and $f_i^2(x)$ at time $T>0$. Let $v_i(x,t)=(u_i^1-u_i^2)(x,t)$. Define the spatial L^2 norm of $v_i(x,t)$ at time t by

$$\| v(t) \| = \left\{ \int_\Omega v_i(x,t)v_i(x,t)dx \right\}^{1/2}. \tag{39}$$

Knops and Payne [18] show that if $u_i^1(x,t)\in\mathcal{P}$ and $u_i^2(x,t)\in\mathcal{Q}$, and if $F(t)=\| v(t) \|^2$,

$$\begin{aligned} F(t)F''(t)-\{F'(t)\}^2 &\ge 2\nu^{-1}(P^2+1)F(t)F'(t) - Q^2\{2\nu^{-2}(P^2+1)+1\}F^2(t) \\ &= -a_1F(t)F'(t)-a_2F^2(t). \end{aligned} \tag{40}$$

Hence, with

$$\begin{aligned} &c=-a_1=2\nu^{-1}(P^2+1), \\ &\mu(t)=(e^{ct}-1)(e^{cT}-1)^{-1}, \\ &w_i(x,t)=e^{-mt}v_i(x,t), \qquad m=-a_2/2a_1, \qquad 0\le t\le T, \end{aligned} \tag{41}$$

it follows from (40) and (26) that

$$\| w(t) \| \le \| w(0) \|^{1-\mu(t)} \| w(T) \|^{\mu(t)}, \qquad 0\le t\le T. \tag{42}$$

Applying Theorem 1 to the Knops–Payne result (42), we have the following corollary.

THEOREM 5 (corollary). *For the given positive ϵ, M, with $\epsilon<M$, let $w_i(x,t)$ in* (41) *satisfy* $\| w(0) \|\le M$, $\| w(T) \|\le\epsilon$, *and let $\mu(t)$ be as in* (41). *Then*

$$\| w(t) \| \le M^{1-\mu(t)}\epsilon^{\mu(t)}, \qquad 0\le t\le T. \tag{43}$$

If, in addition, $\| w(s)-w(0) \|\le K\epsilon$, *with known K, $0<K<M/\epsilon$, and known $s>0$ such that $\mu(s)>\mu^*\equiv\log\{M/(M-K\epsilon)\}/\log(M/\epsilon)$, then*

$$\| w(t) \| \le \Gamma^{1-\mu(t)}\epsilon, \qquad 0\le t\le T, \tag{44}$$

where $\Gamma<M/\epsilon$ is the unique root of $x-x^{1-\mu(s)}-K=0$. Moreover, $\Gamma\ll M/\epsilon$ if $\mu(s)\gg\mu^*$.

For large c, the rapid exponential decay of $\mu(t)$ as t decreases from $t = T$ makes it unlikely that the most general solutions satisfying the constraints (37) or (38) can be continued very far into the past. However, it may be possible to continue solutions that have evolved slowly near $t = 0$. Consider the following example. Let $P = 1,\ \nu = 10^{-1},\ T = 0.25,\ M = 20$, and $\epsilon = 10^{-6}$. Then, $c = 40$ and $\mu(t) = \{e^{40t} - 1\}\{e^{10} - 1\}^{-1},\ 0 \le t \le 0.25$. In particular, $\mu(T/2) = 6.693 \times 10^{-3}$. Consequently, (43) gives

$$\| w(T/2) \| \le 19.603 \times 0.911681 = 17.872 \tag{45}$$

On the other hand, suppose that the solutions to be reconstructed are known to have evolved slowly enough near $t = 0$ that with $s = 0.01T$ and $K = 10$, we have $\| w(s) - w(0) \| \le K\epsilon$. Then $\mu(s) = 4.775 \times 10^{-6}$, while $\mu^* = 2.974 \times 10^{-8}$. Thus, $\{\mu(s)/\mu^*\} = 160.55$. From $\Gamma \log \Gamma \approx K/\mu(s)$, we find $\Gamma \approx 173,600$ and $M/\epsilon = 115\Gamma$. From (44), we get

$$\| w(T/2) \| \le 160,134 \times 10^{-6} = 0.16. \tag{46}$$

Thus, the difference between any two solutions satisfying the SECB constraint is over one hundred times smaller at $t = T/2$ than it is in the more general case of (45).

6. Holomorphic semigroups and evolution equations. Let X be a complex Banach space, let A be a closed linear operator with domain D_A dense in X, and consider the evolution equation $u_t = -Au,\quad t > 0$, for the X-valued function $u(t)$. We assume that $-A$ generates a holomorphic semigroup e^{-tA} in an open sector of the complex t-plane, $\Sigma_\phi = \{\text{Re } t > 0,\ \ |\text{Arg } t| < \phi\}$, for some fixed $\phi,\ \ 0 < \phi \le \pi/2$. Moreover, for any $0 < \sigma < \phi,\ \ e^{-tA}$ is strongly continuous at $t = 0$ within $\Sigma_{\phi-\sigma}$, reduces to the identity operator at $t = 0$, and satisfies $\| e^{-tA} \| \le B_\sigma < \infty$ for $t \in \overline{\Sigma}_{\phi-\sigma}$. Thus, e^{-tA} is a bounded holomorphic semigroup as defined in [15].

Parabolic initial boundary value problems constitute the best-known area of application of holomorphic semigroups. We briefly sketch this connection below, and refer the reader to [12] and [28] for a complete treatment. Less well known are applications to a wide class of nonparabolic equations, typically involving nonlocal partial differential operators, that are obtained by "subordination" in well-posed Cauchy problems [4], [11], [8]. This class of problems, mentioned in section 6.1, is drawing increasing interest from physical scientists working in certain areas of fractal analysis.

Let Ω be a bounded domain in R^n with a sufficiently smooth boundary $\partial\Omega$. For $x \in R^n$, let $A(x, D) = \sum_{|\alpha| \le 2m} a_\alpha(x) D^\alpha$ be a linear partial differential operator with coefficients $a_\alpha(x)$ continuous in the closure of Ω. If $A(x, D)$ is strongly elliptic, and zero Dirichlet data are given on $\partial\Omega$, a closed linear operator A in $L^2(\Omega)$, with dense domain $D_A = H^{2m}(\Omega) \cap H_0^m(\Omega)$, can be defined by

$$(Au)(x) = A(x, D)u(x), \qquad u \in D_A. \tag{47}$$

Moreover, as shown in [12], [28], for some $k \ge 0$ the linear operator $-(A+kI)$ generates a bounded holomorphic semigroup in $L^2(\Omega)$. If $A(x, D)$ is a symmetric differential operator, then $A + kI$ is self-adjoint, and we may choose $\phi = \pi/2$ in Σ_ϕ.

More general boundary conditions can be handled and parabolic equations of order $2m$ can be considered in $L^p(\Omega),\ \ 1 \le p < \infty$. Let $H^{j,p}(\Omega)$ denote the Sobolev space of functions in $L^p(\Omega)$ whose weak derivatives of order less than or equal to j exist and belong to $L^p(\Omega)$. Let $\{B_j\}_{j=1}^m$ be m boundary operators of respective orders $m_j < 2m$, given by

$$B_j(x,D) = \sum_{|\alpha| \le m_j} b_\alpha^j(x) D^\alpha, \tag{48}$$

and consider the boundary value problem

$$\begin{aligned} &A(x,D)u = g, \qquad x \in \Omega, \\ &B_j(x,D)u = 0, \qquad x \in \partial\Omega, \quad 1 \le j \le m. \end{aligned} \tag{49}$$

A closed linear operator A with dense domain $D_A = H^{2m,p}(\Omega; \{B_j\})$, consisting of the closure in $H^{2m,p}(\Omega)$ of the set of functions $u \in C^{2m}(\overline{\Omega})$ that satisfy the boundary conditions in (49), can be defined via

$$(Au)(x) = A(x,D)u, \qquad u \in D_A. \tag{50}$$

If the system B_j is *normal*, and satisfies further *complementary* conditions, and if $A(x,D)$ is strongly elliptic, one obtains a *regular* elliptic boundary value problem, $(A, \{B_j\}, \Omega)$, such that for some $k \ge 0$, the linear operator $-(A + kI)$ generates a bounded holomorphic semigroup in $L^p(\Omega)$. See [12], [28].

6.1. Subordinated semigroups. Let $H(y)$ denote the Heaviside unit step function, and consider the family $p_y(t)$ given by

$$p_y(t) = \frac{tH(y)e^{-t^2/4y}}{\sqrt{4\pi y^3}}, \qquad t > 0. \tag{51}$$

For each fixed $t > 0$, $p_y(t)$ is a probability density function on $y \ge 0$, and $p_y(t)$ tends to the Dirac δ-function $\delta(y)$ as $t \downarrow 0$. Moreover, if $*$ denotes convolution with respect to y, then $p_y(t) * p_y(s) = p_y(t+s)$, for $s,\ t \ge 0$. The Laplace transform with respect to y of $p_y(t)$ is given by

$$\mathcal{L}\{p_y(t)\} \equiv \int_0^\infty e^{-yz} p_y(t)\,dy = e^{-t\sqrt{z}}, \qquad \text{Re } z > 0. \tag{52}$$

The "inverse Gaussian" family in (51) is just one example of an *infinitely divisible* family of probability density functions on the half-line $y \ge 0$, [11].

Let $T(t) = e^{-tA},\ t \ge 0$, be a uniformly bounded, not necessarily holomorphic, C_0 semigroup on a complex Banach space X. Using (51), one may construct a new C_0 semigroup $U(t)$ on X, with $\| U(t) \| \le \| T(t) \| \le B < \infty,\ \ t \ge 0$, by means of

$$U(0) = I, \quad U(t)g = \int_0^\infty p_y(t) T(y) g \; dy, \qquad t > 0, \quad g \in X. \tag{53}$$

Indeed, it turns out that $U(t) = e^{-tA^{1/2}}$ and that $U(t)$ can be extended to a *bounded holomorphic semigroup* in some sector Σ_ω.

The construction in (53) amounts to randomization of the time variable t in the original semigroup $T(t)$. A wide variety of infinitely divisible families $q_y(t)$ may be used in (53). The new semigroup $U(t)$ is said to be "subordinated" to $T(t)$ through the "directing process" $q_y(t)$ [11]. This concept originated in [4] and was subsequently refined into a functional calculus in [26], [23], and [3]. The observation that $U(t)$ is holomorphic whenever the directing process $q_y(t) = \mathcal{L}^{-1}\{e^{-tz^\alpha}\},\ \ 0 < \alpha < 1$, was made in [29]. In that case, $U(t) = e^{-tA^\alpha}$. Subordinated processes and fractional

differential operators are of interest in polymer science [9], while diffusion equations with fractional Laplacians play a role in image deblurring [6]. Further applications are discussed in [5, pp. 140–156] and [11].

An arbitrary infinitely divisible family $q_y(t)$ on $y \geq 0$ can be characterized in terms of its Laplace transform [11]. We have $\mathcal{L}\{q_y(t)\} = e^{-t\psi(z)},\ t \geq 0$, where the *exponent* $\psi(z)$ is holomorphic for Re $z > 0$ and continuous for Re $z \geq 0$, with Re $\psi(z) \geq 0$. Moreover, $\psi(0) = 0$, and $\psi'(x)$ is completely monotone for $x > 0$. In [8], the results of [29] are extended. A necessary and sufficient condition on $q_y(t)$ is given, in order that the subordinated semigroup $U(t) = e^{-t\psi(A)}$ be holomorphic on X, whenever $T(t)$ is C_0 and uniformly bounded on X. In addition, a *necessary* condition on the exponent $\psi(z)$ is obtained for that to be the case. In [13], a *sufficient* condition on $\psi(z)$ is given that ensures analyticity of $U(t)$. As a consequence of [29], [8], and [13], a rich class of exponents $\psi(z)$ is known, with the property that $-\psi(A)$ generates a bounded holomorphic semigroup on X whenever $-A$ generates a uniformly bounded C_0 semigroup on X. As one example, consider the symmetric hyperbolic system,

$$\begin{aligned} u_t &= \sum_{i=1}^{n} a_i(x) u_{x_i} + b(x)u, \qquad x \in R^n, t > 0, \\ u(x,0) &= f(x), \end{aligned} \tag{54}$$

where $u(x)$ is an N-component vector, $a_i(x),\ b(x)$ are $N \times N$ matrices with boundedly differentiable entries on R^n, and $a_i(x)$ is Hermitian. The differential operator on the right-hand side of (54) can be extended into a closed densely defined linear operator $-A$ in $L^2(R^n)^N$. As shown in [28], for some $k \geq 0,\ -(A+kI)$ generates a contraction semigroup on $L^2(R^n)^N$. It follows from [29], [8] that if

$$\psi_1(A) = (A + kI)^{\alpha},\ 0 < \alpha < 1, \qquad \psi_2(A) = \mathrm{Log}\{A + (k+1)I\}, \tag{55}$$

then each of $-\psi_1(A), -\psi_2(A)$, generates a holomorphic semigroup on $L^2(R^n)^N$. If $\{\alpha_n\}_{n=1}^{\infty}$ and $\{a_n\}_{n=1}^{\infty}$ are any two sequences satisfying $a_n \geq 0,\ \ a_1 > 0,\ \ 1 > \alpha_1 > \alpha_2 > \cdots > \alpha_n > \cdots > 0,\ \ \sum_{n=1}^{\infty} a_n/\alpha_n < \infty$, and if

$$\psi_3(A) = \sum_{n=1}^{\infty} a_n (A + kI)^{\alpha_n}, \tag{56}$$

it follows from [13] that $-\psi_3(A)$ generates a holomorphic semigroup on $L^2(R^n)^N$. None of the $\psi_i(A),\ i = 1, 3$, are elliptic operators when $-A$ is the differential operator on the right-hand side of (54). This shows that holomorphic semigroup theory encompasses a class of initial value problems in partial differential equations that is considerably wider than the class of parabolic problems.

7. Logarithmic convexity and holomorphic semigroups. In Banach space, approaches different from those used in sections 3–5 appear necessary to obtain logarithmic convexity inequalities. Following the basic work in [19], further convexity results were obtained in [1], [12], and [22]. Theorems 6 and 7 below are a reformulation of results originating with these authors.

For any $a \geq 0$, and $0 < \xi \leq 1$, let $S(a, \xi)$ be the set in the complex τ-plane given by

$$S(a,\xi) = \{\tau = t + is;\ \ t \geq a;\ \ |s| \leq (t-a)\tan(\pi\xi/2)\}. \tag{57}$$

Let $T > 0$. Then, $S(T,\xi) \subset S(0,\xi)$. Let $G(T,\xi) = S(0,\xi) \setminus S(T,\xi)$, and let Λ_L, Λ_R be, respectively, the left and right boundary arcs of $G(T,\xi)$. Let $\omega_\xi(t,s)$ be the unique bounded continuous function on $\overline{G}(T,\xi)$ which is harmonic in the interior of $G(T,\xi)$, equals zero on Λ_L, and equals one on Λ_R. Let $\mu_\xi(t) = \omega_\xi(t,0)$, $0 \le t \le T$.

LEMMA 2. *$\mu_1(t) = t/T$, and, if $0 < \xi < \eta \le 1$, $\mu_\xi(t) < \mu_\eta(t)$, $0 < t < T$.*

Proof. Let $H(\xi,\eta) = S(0,\eta) \setminus S(T,\xi)$. Then $G(T,\xi) \subset H(\xi,\eta)$, and $G(T,\eta) \subset H(\xi,\eta)$. Let Λ'_L be the left boundary arc of $H(\xi,\eta)$, and let Λ'_R be the right boundary arc of $G(T,\eta)$. Let $\tilde{\omega}(t,s)$ be the unique bounded continuous function on $\overline{H}(\xi,\eta)$ which is harmonic in the interior of $H(\xi,\eta)$, equals zero on Λ'_L, and equals one on Λ_R. The harmonic function $\tilde{\omega} - \omega_\xi$ in $G(T,\xi)$ has value zero on Λ_R, is nonnegative on Λ_L and hence must be strictly positive in the interior of $G(T,\xi)$. Therefore $\mu_\xi(t) < \tilde{\omega}(t,0)$, $0 < t < T$. A similar argument, applied to the harmonic function $\omega_\eta - \tilde{\omega}$ in $G(T,\eta)$, shows that $\mu_\eta(t) > \tilde{\omega}(t,0)$, $0 < t < T$. Finally, if $\xi = 1$, then $\overline{G}(T,1)$ is the vertical strip $0 \le \operatorname{Re}\tau \le T$, and $\omega_1(t,s) = t/T$. □

We now consider the evolution equation $u_t = -Au$, $t > 0$, in a complex Banach space X with norm $\| \quad \|$, under the assumption that $-A$ generates a bounded holomorphic semigroup in an open sector Σ_ϕ in the complex $\tau = t + is$ plane. With $0 < \alpha\pi/2 < \phi \le \pi/2$, let $S(0,\alpha)$, defined in (57), be a closed subsector of Σ_ϕ, and let $\| e^{-\tau A} \| \le B_\alpha < \infty$, $\tau \in S(0,\alpha)$. Introduce the equivalent norm $\| \quad \|_\alpha$ on X defined by

$$\| x \|_\alpha \equiv \sup_{\tau \in S(0,\alpha)} \| e^{-\tau A} x \|, \qquad x \in X. \tag{58}$$

Then, as is easily verified,

$$\| x \| \le \| x \|_\alpha \le B_\alpha \| x \|, \quad x \in X, \qquad \| e^{-\tau A} \|_\alpha \le 1, \quad \tau \in S(0,\alpha). \tag{59}$$

THEOREM 6. *Let X be a complex Banach space with norm $\| \quad \|$, let $u(t)$ be a solution of $u_t = -Au$, $0 < t \le T$, where $-A$ generates a bounded holomorphic semigroup on X. Then, with $\| \quad \|_\alpha$ as in (58) and $\mu_\alpha(t)$ as in Lemma 2,*

$$\| u(t) \|_\alpha \le \| u(0) \|_\alpha^{1-\mu_\alpha(t)} \| u(T) \|_\alpha^{\mu_\alpha(t)}, \qquad 0 \le t \le T, \tag{60}$$

and

$$\| u(t) \| \le B_\alpha \| u(0) \|^{1-\mu_\alpha(t)} \| u(T) \|^{\mu_\alpha(t)}, \qquad 0 \le t \le T. \tag{61}$$

Proof. Let l be a linear functional on X with $|l|_\alpha = 1$, where $| \ |_\alpha$ denotes the norm on X^* corresponding to the norm $\| \quad \|_\alpha$ on X. Let $h(\tau) = l(e^{-\tau A} u(0))$ for $\tau \in S(0,\alpha)$. We have that $h(\tau)$ is continuous and bounded on $S(0,\alpha)$, with $|h(\tau)| \le \| u(0) \|_\alpha$, and $h(\tau)$ is holomorphic in the interior of $S(0,\alpha)$. The same is true for $h(\tau)$ in $S(T,\alpha)$, with $|h(\tau)| \le \| u(T) \|_\alpha$. This follows from $e^{-\tau A} = e^{-(\tau - T)A} e^{-TA}$ for $\tau \in S(T,\alpha)$. Let $G(T,\alpha)$ and $\omega_\alpha(t,s)$ be as defined above, and consider the function $v(t,s)$ in $G(T,\alpha)$ where

$$v(t,s) = \log|h(\tau)| - \omega_\alpha(t,s) \log \| u(T) \|_\alpha + (\omega_\alpha(t,s) - 1) \log \| u(0) \|_\alpha . \tag{62}$$

The function $v(t,s)$ is upper semicontinuous and bounded above on $G(T,\alpha)$, subharmonic in the interior of $G(T,\alpha)$, and nonpositive on the left and right boundary arcs of $G(T,\alpha)$. Therefore $v(t,s) \le 0$ on $\overline{G}(T,\alpha)$. Using

$$\| u(\tau) \|_\alpha = \sup_{l \in X^*, \ |l|_\alpha = 1} |h(\tau)|, \tag{63}$$

we obtain

$$\| u(\tau) \|_\alpha \leq \| u(0) \|_\alpha^{1-\omega_\alpha(\tau)} \| u(T) \|_\alpha^{\omega_\alpha(\tau)}, \qquad \tau \in \overline{G}(T,\alpha), \tag{64}$$

which implies, on using (59),

$$\| u(\tau) \| \leq B_\alpha \, \| u(0) \|^{1-\omega_\alpha(\tau)} \| u(T) \|^{\omega_\alpha(\tau)}, \qquad \tau \in \overline{G}(T,\alpha). \tag{65}$$

Finally, (60), (61), follow from the above on putting $\tau = t$. □

Remark 2. The inequality (61) follows from (60) but not vice versa. From Lemma 2, we see that $\mu_\alpha(t)$ is sublinear in t, and this sublinearity becomes more severe as α becomes smaller. The choice of α depends on the spectrum of the spatial operator A. Since $-A$ generates a holomorphic semigroup in the open sector Σ_ϕ, the spectrum of A must be contained in the closed sector $Arg\,|z| \leq \beta = \pi/2 - \phi$ in the right half-plane. As β increases, ϕ, and hence α, must decrease. Theorem 6 does not yield the explicit dependence of $\mu_\alpha(t)$ on t, which is necessary for applying the SECB constraint. The next result is more useful in that regard.

THEOREM 7. *With $u(t)$ and α as in Theorem* 6, *let* $0 < \sigma < \alpha < 1$, *and let*

$$\begin{aligned} &\lambda = \inf_{0 \leq \theta \leq \pi/2} \{\cos\sigma\theta \, [1 - \tan\sigma\theta / \tan(\alpha\pi/2)] \, / (\cos\theta)^\sigma\}, \\ &\rho_\sigma(t) = (\lambda t/T)^{1/\sigma}, \qquad 0 \leq t \leq T. \end{aligned} \tag{66}$$

Then,

$$\| u(t) \|_\alpha \leq \| u(0) \|_\alpha^{1-\rho_\sigma(t)} \| u(T) \|_\alpha^{\rho_\sigma(t)}, \qquad 0 \leq t \leq T, \tag{67}$$

and

$$\| u(t) \| \leq B_\alpha \, \| u(0) \|^{1-\rho_\sigma(t)} \| u(T) \|^{\rho_\sigma(t)}, \qquad 0 \leq t \leq T. \tag{68}$$

Proof. Note that λ in (66) satisfies $0 < \lambda < 1$ and may be found graphically given α and σ. Let $Y > 0$, let l be a linear functional on X with $|l|_\alpha = 1$, and let $h(\tau) = l(e^{-\tau A} u(0))$ for $\tau \in S(0,\alpha)$. As in Theorem 6, $h(\tau)$ is continuous and bounded on $S(0,\alpha)$ (resp., $S(Y,\alpha)$) and holomorphic in its interior, with $|h(\tau)| \leq \| u(0) \|_\alpha$, (resp., $|h(\tau)| \leq \| u(Y) \|_\alpha$). Let $0 < \sigma < \alpha$, let V be the vertical strip $0 \leq \mathrm{Re}\ \tau \leq Y$, and consider the function $\psi(\tau) = h(\tau^\sigma)$ for $\tau \in V$. We have that $\psi(\tau)$ is continuous and bounded on V, holomorphic in its interior, with $|\psi(\tau)| \leq \| u(0) \|_\alpha$. A more precise estimate for $|\psi(\tau)|$ on the line $\mathrm{Re}\ \tau = Y$ will now be obtained. We first show that with λ as in (66),

$$\mathrm{Re}\ \tau = Y \Longrightarrow \tau^\sigma \in S(\lambda Y^\sigma, \alpha). \tag{69}$$

Indeed, with $\tau = Y + is = re^{i\theta}$, $0 \leq |\theta| < \pi/2$, we have $\tau^\sigma = r^\sigma(\cos\sigma\theta + i \, \sin\sigma\theta)$, and $Y = r\cos\theta$. Therefore, $\tau^\sigma \in S(\lambda Y^\sigma, \alpha)$ if and only if

$$r^\sigma |\sin\sigma\theta| \leq \{r^\sigma \cos\sigma\theta - \lambda(r\cos\theta)^\sigma\} \tan(\alpha\pi/2), \qquad 0 \leq |\theta| < \pi/2, \tag{70}$$

i.e., if and only if $\forall\ 0 \leq \theta < \pi/2$, we have

$$\lambda \leq \cos\sigma\theta \, \{1 - \tan\sigma\theta / \tan(\alpha\pi/2)\} \, / (\cos\theta)^\sigma. \tag{71}$$

But this is guaranteed from the definition of λ. It follows that

$$|\psi(\tau)| \leq \| u(0) \|_\alpha, \quad \text{Re } \tau = 0, \qquad |\psi(\tau)| \leq \| u(\lambda Y^\sigma) \|_\alpha, \quad \text{Re } \tau = Y. \tag{72}$$

We may now apply the "three lines theorem," [27, p. 244], to $\psi(\tau)$ in the strip V and conclude that

$$|\psi(y)| \leq \| u(0) \|_\alpha^{1-y/Y} \| u(\lambda Y^\sigma) \|_\alpha^{y/Y}, \qquad 0 \leq y \leq Y. \tag{73}$$

Using (63), we obtain

$$\| u(y^\sigma) \|_\alpha \leq \| u(0) \|_\alpha^{1-y/Y} \| u(\lambda Y^\sigma) \|_\alpha^{y/Y}, \qquad 0 \leq y \leq Y. \tag{74}$$

Putting $t = y^\sigma,\ T = \lambda Y^\sigma,\ \rho_\sigma(t) = (\lambda t/T)^{1/\sigma}$ in (74) gives

$$\| u(t) \|_\alpha \leq \| u(0) \|_\alpha^{1-\rho_\sigma(t)} \| u(T) \|_\alpha^{\rho_\sigma(t)}, \qquad 0 \leq t \leq T/\lambda. \tag{75}$$

Since $T/\lambda > T$, (75) implies (67) which implies (68). □

Remark 3. When X is a Hilbert space, $\rho_\sigma(t)$ in (66) may be viewed as expressing the penalty for non-self-adjointness in the spatial operator A. When A is self-adjoint, we have $\rho(t) = t/T$. If the spectrum of A leaves the nonnegative real axis and expands into the sector $Arg\ |z| \leq \pi/2 - \phi$, $\rho_\sigma(t)$ decays to zero faster than t/T, through the exponent $1/\sigma$. It is remarkable that (66) actually holds in any complex Banach space X. The next theorem summarizes the main results of this section.

THEOREM 8 (corollary). *Let X be a complex Banach space with norm $\| \quad \|$. Let $-A$ generate a holomorphic semigroup $e^{-\tau A}$ on X, satisfying $\| e^{-\tau A} \| \leq B_\alpha < \infty$, in a closed sector $|Arg\ \tau| \leq \alpha\pi/2$ of the complex $\tau = t + is$ plane, for suitable α with $0 < \alpha < 1$. Let $\| \quad \|_\alpha$ be the equivalent norm on X defined in* (58), (59). *Let $0 < \sigma < \alpha$, and let λ and $\rho_\sigma(t)$ be as in* (66). *For given ϵ, M, with $\epsilon < M$, let $f \in X$ be given data at time $T > 0$, and let $u_i(t)$, $i = 1, 2$, be two solutions of $u_t = -Au + g(t)$, $0 < t \leq T$, with $\| u_i(T) - f \| \leq \epsilon/B_\alpha$, and $\| u_i(0) \| \leq M/B_\alpha$. Finally, let $w(t) = u_1(t) - u_2(t)$. Then*

$$\| w(t) \| \leq \| w(t) \|_\alpha \leq 2M^{1-\rho_\sigma(t)} \epsilon^{\rho_\sigma(t)}, \qquad 0 \leq t \leq T. \tag{76}$$

If, in addition, $\| u_i(s) - u_i(0) \| \leq K\epsilon/B_\alpha$, $i = 1, 2$, with known K, $0 < K < M/\epsilon$, and known $s > 0$ such that $\rho_\sigma(s) > \mu^*$, where μ^* is defined in (8), *then*

$$\| w(t) \| \leq \| w(t) \|_\alpha \leq 2\Gamma^{1-\rho_\sigma(t)} \epsilon, \qquad 0 \leq t \leq T, \tag{77}$$

where $\Gamma < M/\epsilon$ is the unique root of $x - K - x^{1-\rho_\sigma(s)} = 0$. Moreover, $\Gamma \ll M/\epsilon$ if $\rho_\sigma(s) \gg \mu^*$.

Proof. From (59), we have $\| w(0) \|_\alpha \leq 2M$, $\| w(T) \|_\alpha \leq 2\epsilon$. Hence, (76) follows from (67). Likewise, $\| w(s) - w(0) \|_\alpha \leq 2K\epsilon$. Applying Theorem 1 with the $\| \quad \|_\alpha$ norm on X, we obtain (77) from (67). □

8. An example. In the Navier–Stokes equations, where the Hölder exponent $\mu(t)$ in (41) depends on $1/\nu$, it is not known whether or not there can be equality in the Knops–Payne inequality (42). However, the following example demonstrates that rapid exponential decay in $\mu(t)$ can be realized in quite simple problems. With positive constants $a,\ c,\ Q$, consider the 1-D parabolic initial value problem in $L^2(0, \pi)$,

$$\begin{aligned} &u_t = ae^{ct} u_{xx}, \qquad 0 < x < \pi, \quad t > 0, \\ &u(0, t) = u(\pi, t) = 0, \qquad t \geq 0, \\ &u(x, 0) = Q \sin mx, \qquad 0 \leq x \leq \pi. \end{aligned} \tag{78}$$

The unique solution of (78) is

$$u(x,t) = Q\ e^{-am^2(e^{ct}-1)/c}\ \sin mx, \qquad t \geq 0. \tag{79}$$

Moreover, $u(x,t)$ satisfies

$$\| u(t) \| = \| u(0) \|^{1-\mu(t)} \| u(T) \|^{\mu(t)}, \qquad 0 \leq t \leq T, \tag{80}$$

where $\mu(t) = \{e^{ct}-1\}\{e^{cT}-1\}^{-1}$ and $\|\ \ \|$ is the norm on $L^2(0,\pi)$. This shows that Theorem 3 is sharp. By choosing $c > 0$ sufficiently large in (78), we can expect to simulate some of the difficulties that would attend backwards in time continuation in the Navier–Stokes equations.

Let $a = 2 \times 10^{-5}$, let $c = 10$, and, for any positive integer m, let

$$g_m(t) = e^{-am^2(e^{ct}-1)/c}, \qquad t \geq 0. \tag{81}$$

Let $p = \sqrt{2/\pi}$. With $M = 10$, and $\epsilon = 2 \times 10^{-7}$, consider the initial data

$$u(x,0) = p\sqrt{(1-\epsilon^2)/2}\ M \sin 2x + p \sum_{n=1}^{\infty} b_{2n+1} \sin(2n+1)x, \tag{82}$$

where

$$\sum_{n=1}^{\infty} b_{2n+1}^2 = \epsilon^2 M^2/2, \qquad \sum_{n=1}^{\infty} n^q b_{2n+1}^2 = \infty \qquad \forall\ q > 0. \tag{83}$$

Thus, $u(x,0)$ is an L^2 function on $(0,\pi)$ which is not in $H^q(0,\pi)$ for any $q > 0$, and $\| u(0) \| = M/\sqrt{2}$. We may think of the second term in (82) as representing highly localized, nondifferentiable singularities that are superimposed onto the first term. With these initial data in (78), the unique solution is

$$u(x,t) = p\sqrt{(1-\epsilon^2)/2}\ M g_2(t)\ \sin 2x + p \sum_{n=1}^{\infty} b_{2n+1}\ g_{2n+1}(t)\ \sin(2n+1)x. \tag{84}$$

Given an a priori L^2 bound for $u(x,t)$ at $t = 0$, consider recovering the solution (84) on $0 \leq t < 1$, from approximate data $f(x)$ at $t = 1$, with $\| u(1) - f \| \leq \epsilon$. Let

$$\| u(0) \| \leq M = 10 \tag{85}$$

be this prescribed bound, and let the data $f(x)$ at $t = 1$ be given by

$$f(x) = u(x,1) + p(M/\sqrt{2})\ g_{20}(1)\ \sin 20x. \tag{86}$$

Then

$$\begin{aligned} &\| u(1) - f \| = (M/\sqrt{2}) g_{20}(1) = 1.574 \times 10^{-7} < \epsilon, \\ &\| u(1) - f \| \ / \ \| u(1) \| < g_{20}(1)\{\sqrt{1-\epsilon^2}\ g_2(1)\}^{-1} = 2.66 \times 10^{-8}. \end{aligned} \tag{87}$$

Evidently, the given data $f(x)$ approximates $u(x,1)$ extremely closely in both absolute and relative terms. However, if $v(x,t)$ is the function

$$v(x,t) = u(x,t) + p(M/\sqrt{2})\ g_{20}(t)\ \sin 20x, \qquad 0 \leq t \leq 1, \tag{88}$$

then $v(x,1) = f(x)$, $v(x,t)$ is a solution, and, since $\| v(0) \| = M$, $v(x,t)$ is an equally valid continuation. Noteworthy is the substantial qualitative difference between $u(x,t)$ and $v(x,t)$, which emerges as early as $t = 1/2$, as continuation unfolds backwards from $t = 1$. While $g_2(1) = 0.8384$ and $g_{20}(1) = 2.226 \times 10^{-8}$, we find $g_2(1/2) = 0.9988$ and $g_{20}(1/2) = 0.8888$. Consequently, while the primary component in $u(x,t)$ is the large amplitude $\sin 2x$ oscillation for $0 \le t \le 1$, the $\sin 2x$ and $\sin 20x$ terms have approximately *equal* amplitudes in $v(x,t)$, for $0 \le t \le 1/2$. Clearly, Hölder-continuous data dependence is simply too weak to distinguish $u(x,t)$ from $v(x,t)$ in this example, even though $\| u(1/2) - v(1/2) \| = 6.285$ is roughly the same size as $\| u(0 \|$.

We shall show that an SECB constraint can easily distinguish between $u(x,t)$ and $v(x,t)$, although neither function is differentiable in x at $t = 0$. Indeed, with $K = 35$ and $s = 0.01$, we find

$$\begin{aligned} \| u(s) - u(0) \|^2 &= (1-\epsilon^2)(M^2/2)(1-g_2(s))^2 + \sum_{n=1}^{\infty} b_{2n+1}^2 \, (1 - g_{2n+1}(s))^2, \\ &\le (1-\epsilon^2)(M^2/2)(1-g_2(s))^2 + \epsilon^2 M^2/2, \\ &= (6.115 \times 10^{-6})^2 < K^2 \, \epsilon^2. \end{aligned} \tag{89}$$

On the other hand, with $s = 0.01$,

$$\begin{aligned} \| v(s) - v(0) \|^2 &= \| u(s) - u(0) \|^2 + (M^2/2)(1 - g_{20}(s))^2, \\ &> (M^2/2)(1 - g_{20}(s))^2 = (5.949 \times 10^{-4})^2, \\ &> (2974)^2 \, \epsilon^2. \end{aligned} \tag{90}$$

Therefore, the SECB constraint

$$\| u(0.01) - u(0) \| \le 35 \, \epsilon \tag{91}$$

eliminates $v(x,t)$ in (88) as a possible continuation, while allowing $u(x,t)$ in (84). Here, $\mu(s)/\mu^* = 121$, $\Gamma = 554,235$, and $M/\epsilon = 90\Gamma$. It follows from Theorem 1 and (80) that if $u_1(x,t)$ is any other continuation satisfying (91), then $\| u(t) - u_1(t) \| \le 2\Gamma^{1-\mu(t)}\epsilon$, $0 \le t \le 1$. Hence, $\| u(1/2) - u_1(1/2) \| \le 0.203$, and $\| u(0) - u_1(0) \| \le 0.222$. Since $\| u(1/2) \| > \{(1-\epsilon^2)/2\}^{1/2} \, M g_2(1/2) = 7.063$, and $\| u(0) \| = 7.071$, the maximum L^2 relative errors in approximating $u(x,t)$ at $t = 1/2$ and at $t = 0$, are, respectively, 2.87% and 3.14%. Without the SECB constraint (91), these relative errors are, respectively, 251% and 283%.

REFERENCES

[1] S. Agmon and L. Nirenberg, *Properties of solutions of ordinary differential equations in Banach space*, Comm. Pure Appl. Math., 16 (1963), pp. 121–239.

[2] K. A. Ames and B. Straughan, *Non-standard and Improperly Posed Problems*, Academic Press, New York, 1997.

[3] A. V. Balakrishnan, *An operational calculus for infinitesimal generators of semigroups*, Trans. Amer. Math. Soc., 91 (1959), pp. 330–353.

[4] S. Bochner, *Diffusion equation and stochastic processes*, Proc. Nat. Acad. Sci. USA, 35 (1949), pp. 368–370.

[5] P. L. Butzer and H. Berens, *Semi-Groups of Operators and Approximation*, Springer-Verlag, New York, 1967.

[6] A. S. Carasso, *Overcoming Hölder continuity in ill-posed continuation problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1535–1557.

[7] A. S. Carasso, *Error bounds in nonsmooth image deblurring*, SIAM J. Math. Anal., 28 (1997), pp. 656–668.

[8] A. S. Carasso and T. Kato, *On subordinated holomorphic semigroups*, Trans. Amer. Math. Soc., 327 (1991), pp. 867–877.
[9] J. F. Douglas, *Some applications of fractional calculus to polymer science*, in Advances in Chemical Physics, 102, I. Prigogine and S. Rice, eds., Wiley, New York, 1997.
[10] L. Eldén, *Numerical solution of the sideways heat equation by difference approximation in time*, Inverse Problems, 11 (1995), pp. 913–923.
[11] W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 2, 2nd ed., Wiley, New York, 1971.
[12] A. Friedman, *Partial Differential Equations*, Holt, Rinehart and Winston, New York, 1969.
[13] Y. Fujita, *A sufficient condition for the Carasso-Kato theorem*, Math. Ann., 297 (1993), pp. 335–341.
[14] F. John, *Continuous dependence on data for solutions of partial differential equations with a prescribed bound*, Comm. Pure Appl. Math., 13 (1960), pp. 551–585.
[15] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1980.
[16] R. J. Knops, ed., *Symposium on Non-Well-Posed Problems and Logarithmic Convexity*, Lecture Notes in Math. 316, Springer-Verlag, New York, 1973.
[17] R. J. Knops, *Logarithmic convexity and other techniques applied to problems in continuum mechanics*, in Symposium on Non-Well-Posed Problems and Logarithmic Convexity, R. J. Knops, ed., Lecture Notes in Math. 316, Springer-Verlag, New York, 1973, pp. 41–54.
[18] R. J. Knops and L. E. Payne, *On the stability of solutions of the Navier–Stokes equations backward in time*, Arch. Rational Mech. Anal., 29 (1968), pp. 331–335.
[19] S. G. Kreĭn and O. I. Prozorovskaya, *Analytic semigroups and incorrect problems for evolutionary equations*, Dokl. Akad. Nauk. SSSR, 13 (1960), pp. 277–280.
[20] H. C. Kuhlmann, M. Wanschura and H. J. Rath, *Flow in two-sided lid-driven cavities: non-uniqueness, instabilities, and cellular structures*, J. Fluid Mech., 336 (1997), pp. 267–299.
[21] K. Miller, *Least squares methods for ill-posed problems with a prescribed bound*, SIAM J. Math. Anal., 1 (1970), pp. 52–74.
[22] K. Miller, *Logarithmic convexity results for holomorphic semigroups*, Pacific J. Math., 58 (1975), pp. 549–551.
[23] E. Nelson, *A functional calculus using singular Laplace integrals*, Trans. Amer. Math. Soc., 88 (1958), pp. 400–413.
[24] L. E. Payne, *Improperly Posed Problems in Partial Differential Equations*, CBMS-NSF Reg. Conf. Series in Appl. Math., 22, SIAM, Philadelphia, PA, 1975.
[25] L. E. Payne, *Some remarks on ill-posed problems for viscous fluids*, Internat. J. Engrg. Sci., 30 (1992), pp. 1341–1347.
[26] R. S. Phillips, *On the generation of semigroups of linear operators*, Pacific J. Math., 2 (1952), pp. 343–369.
[27] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1966.
[28] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.
[29] K. Yosida, *Fractional powers of infinitesimal generators and the analyticity of the semigroups generated by them*, Proc. Japan Acad., 36 (1960), pp. 86–89.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 497–515

RIEMANN PROBLEMS WITH A KINK*

HELGE HOLDEN† AND NILS HENRIK RISEBRO‡

Abstract. We study the Riemann problem for isothermal flow of a gas in a thin pipe with a kink in it. This is modeled by a 2×2 system of conservation laws with Dirac measure sink term concentrated at the location of the bends in the pipe. We show that the Riemann problem for this system of equations always has a unique solution, given an extra condition relating the speeds on both sides of the kink. Furthermore, we study the related problem where the flow is perturbed by a continuous addition of momentum at distinct points. Under certain conditions we show that this Riemann problem also has a unique solution.

Key words. Riemann problem, isothermal gas dynamics, nonlinear resonance

AMS subject classifications. 35L65, 45L67, 76N15

PII. S0036141097327033

0. Introduction. We consider the flow of an isothermal gas in a (infinitely) long, thin pipe of constant cross section. If the walls of the pipe have no effect on the flow, and the pipe is straight, this can be modeled by the system of conservation laws [15, p. 56]

$$\rho_t + (\rho v)_x = 0, \qquad (\rho v)_t + (\rho v^2 + \rho)_x = 0. \tag{0.1}$$

Here, $\rho(x,t)$ denotes the density of the gas, and $v(x,t)$ the velocity. The position along the pipe is described by the coordinate x, and t denotes the time variable. These equations describe the conservation of mass and momentum, respectively.

We here discuss the isothermal case rather than the more general case of polytropic gas modeled by replacing the second equation in (0.1) by $(\rho v)_t + (\rho v^2 + \kappa\rho^\gamma)_x = 0$. This equation yields more unwieldy calculations, and thus we focus on the isothermal case here.

In this paper we discuss the situation where the pipe is not straight but has one or several kinks in it. In between these kinks the pipe is straight. Hence the pipe can be described by a polygonal curve, and we ignore gravity. As in the model without kinks, we let $\rho(x,t)$ denote the density of the gas, and $v(x,t)$ its velocity. We now let x be the arc-length parameter along the pipe, or rather the curve describing the pipe. Away from the kinks, conservation of mass and momentum is given by (0.1). It remains to determine the equations holding at the kinks.

Since the cross section of the pipe is assumed to be constant on each side of a kink, conservation of mass reads, as before,

$$\rho_t + (\rho v)_x = 0. \tag{0.2}$$

In general, we cannot assume that ρ and v are continuous at the location of the kink. Since a kink is always located at the same x, which for simplicity we assume to be at

*Received by the editors September 10, 1997; accepted for publication (in revised form) June 4, 1998; published electronically March 19, 1999.
http://www.siam.org/journals/sima/30-3/32703.html
†Department of Mathematical Sciences, Norwegian University of Science and Technology, N–7491 Trondheim, Norway (holden@math.ntnu.no).
‡Department of Mathematics, University of Oslo, P.O. Box 1053, Blindern, N–0316 Oslo, Norway (nilshr@math.uio.no).

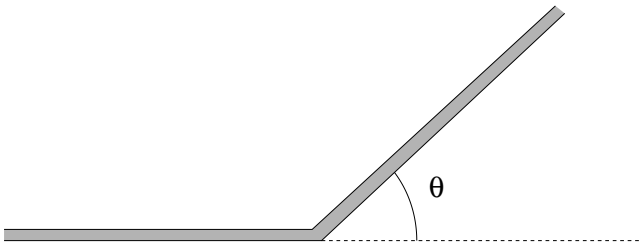


FIG. 1. *The kink.*

$x = 0$, discontinuities at kinks must satisfy a Rankine–Hugoniot condition where the speed of the discontinuity is zero. Hence, from (0.2) we obtain

$$0 = 0\,(\rho_l - \rho_r) = (\rho v)_l - (\rho v)_r,$$

where we have used the notation $f_{l,r} = \lim_{x \to 0^{\mp}} f(x)$. Therefore the product ρv is continuous across kinks. To derive the momentum balance, we again consider a kink located at $x = 0$; the angle of the kink is given by θ (see Figure 1). Since the velocity of the gas is assumed to be parallel to the pipe (except at the kinks), we have

$$\mathbf{v}_l = v_l(1, 0), \qquad \mathbf{v}_r = v_r(\cos\theta, \sin\theta).$$

Consequently, the change in momentum introduced by the kink is given by

$$\rho_r \mathbf{v}_r - \rho_l \mathbf{v}_l = \rho v(\cos\theta - 1, \sin\theta).$$

Therefore, the kink will act as a momentum sink, with a magnitude given by

$$|\rho_r \mathbf{v}_r - \rho_l \mathbf{v}_l| = \rho v \sqrt{2(1 - \cos\theta)}. \tag{0.3}$$

To compensate for the complicated and probably genuinely two-dimensional behavior at the kink, we introduce a multiplicative empirical factor $f \in [0, 1]$. This factor has dimension length(time)$^{-1}$ and is assumed to depend on the properties of the pipe and the gas. For simplicity, we will assume that the pipe is homogeneous, such that f is not dependent on location. Hence, we then arrive at the following model:

$$\begin{aligned} \rho_t + (\rho v)_x &= 0, \\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x &= -f \sum_i k_i \delta_{x_i} \rho v, \end{aligned} \tag{0.4}$$

where δ_{x_i} is the Dirac measure located at the position x_i of the ith kink and k_i is given by the angle of this kink θ_i as

$$k_i = \sqrt{2\,(1 - \cos\theta_i)}.$$

In order to examine one simple consequence of the model (0.4), we imagine a closed piecewise linear pipe approximating a circle of radius r. Assume that we have n equally spaced kinks, each of an angle $\theta = 2\pi/n$. We are interested in what happens for large n; then

$$k_i = \sqrt{2(1 - \cos\,(2\pi/n))} \approx \frac{2\pi}{n}.$$

Also, the distance between adjacent kinks $\Delta x = x_i - x_{i-1} \approx r2\pi/n$. Therefore

$$\sum_i^n \rho v k_i \delta_{x_i} \approx \sum_i^n \rho v \frac{\Delta x}{r} \rightharpoonup \frac{\rho(x,t)v(x,t)}{r},$$

where $\rightharpoonup$ denotes weak convergence. Hence, for a smooth pipe, conservation of momentum is expressed by

$$(\rho v)_t + \left(\rho v^2 + \rho\right)_x = -\rho v f \kappa(x), \tag{0.5}$$

with $\kappa(x)$ denoting the curvature of the pipe at x and $f(x)$ the local empirical bending factor at x.

In our model (0.4) $k_i \in \langle 0, 2\rangle$. Mathematically, however, one can study the model for any value of k_i. For k_i nonnegative we find that (0.4) has a unique solution if k_i is less than 4. In our discussion we are mostly concerned with $k_i = \sqrt{2\,(1-\cos\theta_i)}$.

When one continuously *adds* momentum to the gas at distinct points x_i, we obtain the model (0.4) with k_i negative. The quantity $|fk_i|$ is proportional to the added momentum. For $k_i \geq 2(1-\sqrt{2})$ the equation has a unique solution with the appropriate entropy condition. This case is studied *mutatis mutandis* in section 3.

The model (0.4) also arises as a model for the boundary behavior in an important two-dimensional system of conservation laws. Consider the two-dimensional version of (0.1),

$$\begin{aligned} \rho_t + (\rho v)_x + (\rho u)_y &= 0, \\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x + (\rho v u)_y &= 0, \\ (\rho u)_t + (\rho v u)_x + \left(\rho u^2 + \rho\right)_y &= 0. \end{aligned} \tag{0.6}$$

Here, v and u denote the velocity in the x and y direction, respectively. Let $p(x)$ denote the function

$$p(x) = \begin{cases} 0 & \text{for } x < 0, \\ x\tan\theta & \text{for } x \geq 0, \end{cases}$$

and let Ω denote the set

$$\{(x,y) \mid y > p(x)\}.$$

Then consider (0.6) in Ω with the boundary condition that the velocity is parallel to $\partial\Omega$ at $\partial\Omega$. This system models the isothermal flow across a ramp in two dimensions. If one imposes the initial condition

$$(\rho, u, v)(x,0) = \begin{cases} (\rho_l, u_l, v_l) & \text{for } x < 0, \\ (\rho_r, u_r, v_r) & \text{for } x \geq 0, \end{cases}$$

the solution along the boundary will be given by the solution of the Riemann problem considered here, (1.10). As an application, one could envision using the solution computed in this paper as input at the boundary in a numerical scheme for solving (0.6).

The model (0.5) is an example of a system of conservation laws with source, sometimes referred to as balance equations. The general form of such equations is

$$w_t + f(w)_x = g(x,w), \tag{0.7}$$

where $w = w(x,t)$ is in $\mathbb{R}^n$, and consequently f is a mapping $\mathbb{R}^n \to \mathbb{R}^n$ and g is a mapping $\mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$. Balance equations of this form are used to model a variety of situations. One system of equations, which is somewhat related to the model presented here, is a quasi-one-dimensional model for gas flow in a variable area duct. This model reads

$$\begin{aligned} \rho_t + (\rho v)_x &= -\frac{a'(x)}{a(x)}\rho v, \\ (\rho v)_t + \left(\rho v^2 + p(v)\right)_x &= -\frac{a'(x)}{a(x)}\rho v^2, \\ (\rho E)_t + (\rho E v + p(v)v)_x &= -\frac{a'(x)}{a(x)}\left(\rho E v + p(v)v\right), \end{aligned} \tag{0.8}$$

where ρ and v are as before, and $p(v)$ denotes the pressure and E the total energy of the gas. The cross sectional area of the duct is denoted by $a(x)$. This model (0.8) has been analyzed by Liu in [4, 5, 7, 8]. A scalar version was analyzed by Greenberg et al. [3]. A discussion of multiple steady states in the context of one-dimensional transonic flow can be found in [1]. If the source term g is smooth and bounded, and if the eigenvalues of the Jacobian df are real, distinct, and different from zero, existence of a global (in t) solution was obtained in [4] by a generalization of Glimm's method, and uniqueness and stability were proved by Crasta and Piccoli [6] in the 2×2 case. A numerical method for such systems was analyzed by Glaister [2]. If, however, the eigenvalues may take the value zero, (0.7) is a so-called resonant hyperbolic system. This is the case for the system (0.8), as well as for the system discussed in this paper. Chen and Glimm [11] proved existence of global solutions of the system consisting of the first two equations in (0.8) with a smoothly varying cross section.

The wave structure for resonant hyperbolic systems may be surprisingly complicated; see Isaacson and Temple [12], as well as the above-mentioned works by Liu.

If the cross sectional area of the duct in (0.8) is piecewise constant, the source term becomes a point source similar to the source in (0.4). The Riemann problem for (0.8) with a piecewise constant a was analyzed by Marchesin and Paes-Leme in [9].

In addition, we mention that systems of equations exhibiting nonlinear resonance also occur in models of two- and three-phase flow in porous media; see [10] and [13].

1. The Riemann problem. We consider the following system of equations:

$$\begin{aligned} \rho_t + (\rho v)_x &= 0, \\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x &= -\sum_i f_i k_i \delta_{x_i} \rho v, \\ \rho(x,0) = \rho_0(x), &\qquad v(x,0) = v_0(x), \end{aligned} \tag{1.1}$$

where the unknowns ρ and v are functions of x and t. The effects of the bends in the pipe are expressed by the term $k_i \delta_{x_i} \rho v$, where δ_{x_i} denotes the unit point mass located at x_i and k_i is given by

$$k_i = k(\theta_i) = \sqrt{2\left(1 - \cos\left(\theta_i\right)\right)}\,, \tag{1.2}$$

where θ_i is the angle of the bend located at x_i. Even in the absence of source terms, (1.1) generally develops discontinuities, so we seek weak solutions. By definition these

satisfy

(1.3)
$$\iint \rho\varphi_t + \rho v\varphi_x \, dxdt + \int \rho_0(x)\varphi(x,0)\, dx = 0,$$
$$\iint \rho v\varphi_t + \left(\rho v^2 + \rho\right)\varphi_x \, dxdt + \int \rho_0(x)v_0(x)\varphi(x,0)\, dx = \int \sum_i f_i k_i \varphi\left(x_i,t\right)\rho\left(x_i,t\right)v\left(x_i,t\right)\, dt.$$

If there are no source terms, (1.1) is a strictly hyperbolic conservation law, with eigenvalues

$$\lambda_1 = v - 1 \text{ and } \lambda_2 = v + 1. \tag{1.4}$$

The corresponding eigenvectors are

$$e_1 = (1, v-1) \text{ and } e_2 = (1, v+1). \tag{1.5}$$

When solving the Riemann problem we are interested in those states (ρ, v) that can be joined to a given state (ρ_0, v_0) by a simple wave, i.e., either a shock wave or a rarefaction wave. The shock waves satisfy the Lax entropy condition. For a definition of these concepts, see [14, 15]. Through each (ρ_0, v_0) there are two curves, $C_1(\rho_0, v_0)$ and $C_2(\rho_0, v_0)$, of such states. If (ρ, v) is on $C_1(\rho_0, v_0)$, then the initial value problem

$$\begin{gathered} \rho_t + (\rho v)_x = 0, \\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x = 0, \\ \rho(x,0) = \begin{cases} \rho_0 & \text{for } x < 0, \\ \rho & \text{for } x > 0, \end{cases} \qquad v(x,0) = \begin{cases} v_0 & \text{for } x < 0, \\ v & \text{for } x > 0 \end{cases} \end{gathered} \tag{1.6}$$

is solved by a slow wave (also denoted as a one-wave). This wave is a (one-)shock wave if $\rho > \rho_0$ and a (one-)rarefaction wave if $\rho < \rho_0$. The curve C_1 is given by the following expression [15, pp. 71ff and p. 84f]:

$$C_1 \;:\; v_1(\rho; \rho_0, v_0) = \begin{cases} v_0 - \ln\left(\frac{\rho}{\rho_0}\right) & \text{for } \rho < \rho_0, \\ v_0 - \frac{\rho - \rho_0}{\sqrt{\rho\rho_0}} & \text{for } \rho > \rho_0. \end{cases} \tag{1.7}$$

Similarly, if (ρ, v) is on $C_2(\rho_0, v_0)$, then the initial value problem (1.6) is solved by either a two-shock wave or a two-rarefaction wave. The curve C_2 is given by [15, pp. 71ff and p. 84f]

$$C_2 \;:\; v_2(\rho; \rho_0, v_0) = \begin{cases} v_0 + \frac{\rho - \rho_0}{\sqrt{\rho\rho_0}} & \text{for } \rho < \rho_0, \\ v_0 + \ln\left(\frac{\rho}{\rho_0}\right) & \text{for } \rho > \rho_0, \end{cases} \tag{1.8}$$

with two-shock waves corresponding to $\rho < \rho_0$ and two-rarefaction waves corresponding to $\rho > \rho_0$. Whenever convenient we will denote the shock and rarefaction part of C_i by S_i and R_i, respectively. Furthermore, when we consider the *right state* R as fixed, we denote the corresponding wave curves by C_i^-, etc. Note that C_1

is given by a decreasing convex function of $v_1(\rho)$, such that $\lim_{\rho\to 0+} v_1 = \infty$ and $\lim_{\rho\to\infty} v_1 = -\infty$, and that C_2 is given by an increasing concave function $v_2(\rho)$, such that $\lim_{\rho\to 0+} v_2 = -\infty$ and $\lim_{\rho\to\infty} v_2 = -\infty$. If the source terms in (1.1) are absent, the unique solution of the Riemann problem is ensured.

The speed of shock waves is given by the Rankine–Hugoniot condition and may be computed from the first equation in (1.1):

$$s_j = \frac{\rho v_j - \rho_0 v_0}{\rho - \rho_0} = v_0 + (-1)^j \sqrt{\frac{\rho}{\rho_0}} = v_j + (-1)^j \sqrt{\frac{\rho_0}{\rho}}. \tag{1.9}$$

Note that both shock speeds and the speed of rarefaction waves can be both positive and negative. In our construction of the solution of the Riemann problem we will need certain points on the shock curves. If (ρ, v) is a given state, we denote the point on the shock curve that can be connected with a shock of zero speed by $Z(\rho, v)$. From expressions (1.7)–(1.9) we find that

$$Z(\rho, v) = \left(\rho v^2, \frac{1}{v}\right);$$

for one-shocks it is defined for $v \geq 1$, while for two-shocks it is defined for $v \in [-1, 0\rangle$. Furthermore, we will need the intersection of $C_1(L)$ and the line $v = -1$, and we denote by $\widehat{L}$ the unique point in $C_1(L) \cap \{v = -1\}$. In addition, we let $\widetilde{L}$ denote the unique intersection of $C_1(L)$ and the line $v = v_c^-$ (see (2.4)).

The Riemann problem for (1.1) is the initial value problem

$$\begin{gathered} \rho_t + (\rho v)_x = 0, \\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x = -k\delta_0 \rho v, \\ \rho(x,0) = \begin{cases} \rho_L & \text{for } x < 0, \\ \rho_R & \text{for } x > 0, \end{cases} \qquad v(x,0) = \begin{cases} v_L & \text{for } x < 0, \\ v_R & \text{for } x > 0, \end{cases} \end{gathered} \tag{1.10}$$

where we have absorbed the empirical factor f in the geometric factor k. We study two distinct but related cases; k positive (where $\theta = \arccos(1-k^2/2)$ denotes the angle of the kink) and k negative. In both cases, $|k| \leq 2$. We seek self-similar solutions to (1.10); that is, $\rho = \rho\,(x/t)$ and $v = v\,(x/t)$. Away from the point $x = 0$, we can use the curves C_1 and C_2 to connect states (ρ, v). We label such connections C-waves. At the point $x = 0$, we will in general have a discontinuity. This discontinuity will satisfy the Rankine–Hugoniot conditions

$$[\![\rho v]\!] = 0, \qquad [\![\rho v^2 + \rho]\!] = -k\rho v, \tag{1.11}$$

where by $[\![\phi]\!]$ we denote the jump in ϕ, i.e., $[\![\phi]\!] = \phi_r - \phi_l$. (Recall that the momentum ρv is continuous across the kink.) Given (ρ_l, v_l), the last equation in (1.11) can be solved for v_r, giving

$$v_r := g_\pm\,(v_l) := \frac{1}{2v_l}\left(\alpha\,(v_l) \pm \sqrt{\alpha^2\,(v_l) - 4v_l^2}\right), \tag{1.12}$$

where

$$\alpha(v) = v^2 - kv + 1. \tag{1.13}$$

We can now use the first equation in (1.11) to calculate ρ_r,

$$\rho_r := \rho_l f_\pm(v_l) := \rho_l \frac{v_l}{g_\pm(v_l)}. \tag{1.14}$$

For any point (ρ_l, v_l) we therefore have two candidates for (ρ_r, v_r). In order to choose between these we impose the extra condition

$$\frac{dv_r}{dv_l} \geq 0; \tag{1.15}$$

an increase in velocity of the gas coming into the kink should result in an increase in outgoing velocity. We will use the same criterion for k negative. The choice (1.15) selects a branch of $g_\pm$ and hence of $f_\pm$. The remaining discussion will depend on the properties of this function determined by the sign of k.

2. The Riemann problem for flow through a kink. We will need detailed properties of the function $g_\pm$ which satisfies

$$vg^2 - \alpha(v)g + v = 0, \quad g = g_\pm, \tag{2.1}$$

and this implies that the solution satisfies

$$v = -g(-g(v)). \tag{2.2}$$

The two solutions fulfill

$$g_\pm(v) = g_\pm(1/v), \quad g_+(v)g_-(v) = 1. \tag{2.3}$$

Define

$$v_c^+ = -g_+(-1), \quad v_c^- = -g_-(-1). \tag{2.4}$$

We have that $v_c^- \leq 1 \leq v_c^+$, equality holding only if $k = 0$. The symmetry (2.2) implies that $g_\pm(v_c^\pm) = 1$. Furthermore,

$$\begin{aligned} &g'_+(v) \geq 0 \quad \text{for } v \leq -1 \text{ and for } v \geq v_c^+, \\ &g'_-(v) \geq 0 \quad \text{for } v \in \left[-1, v_c^-\right]. \end{aligned}$$

In Figure 2 below we show $g_\pm$ and $f_\pm$ for $k = 0.5$.

This means that for $v_l \in \langle -\infty, -1] \cup [v_c^+, \infty\rangle$, we choose the plus sign in (1.12) and (1.14), while for $v \in [-1, v_c^-]$, we choose the minus sign. There are no solutions in the region $\langle v_c^-, v_c^+\rangle$. In the figure this choice is indicated by solid lines. Therefore let the mapping K be given by

$$K(\rho, v) = \begin{cases} (\rho f_+(v), g_+(v)) & \text{for } v \leq -1 \text{ and for } v \geq v_c^+, \\ (\rho f_-(v), g_-(v)) & \text{for } v \in [-1, v_c^-]. \end{cases} \tag{2.5}$$

In the following we use the term "K-wave," meaning the mapping K. An important property of the K mapping is that it commutes with the stationary shocks.

LEMMA 2.1. *The K mapping commutes with Z, i.e., $Z(K(\rho, v)) = K(Z(\rho, v))$ whenever the two mappings are defined.*

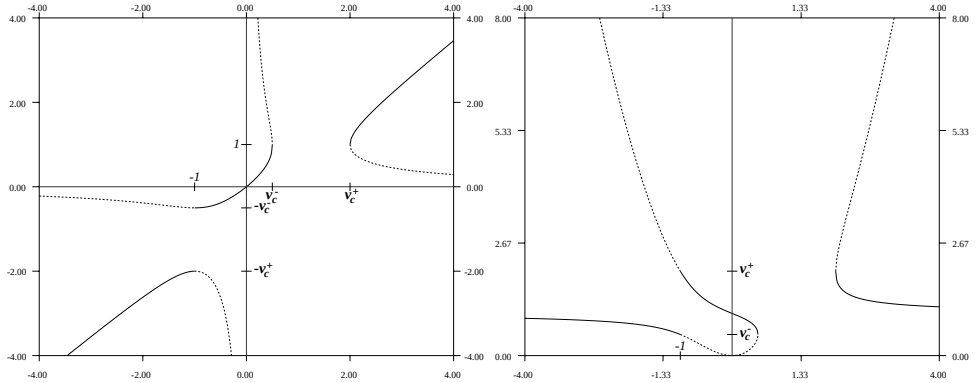


FIG. 2. *The functions $g_\pm$ (left) and $f_\pm$ (right) for $k = 0.5$.*

Proof. Let (ρ, v) be a point with $v \geq v_c^+$. We then compute

$$\begin{aligned} Z\left(K\left(\rho,v\right)\right) &= Z\left(\rho v g_+\left(v\right), g_-\left(v\right)\right) \\ &= \left(\rho v g_+\left(v\right) g_-^2\left(v\right), \frac{1}{g_-\left(v\right)}\right) \\ &= \left(\rho v g_-\left(v\right), g_+\left(v\right)\right) \end{aligned}$$

and

$$\begin{aligned} K\left(Z\left(\rho,v\right)\right) &= K\left(\rho v^2, \frac{1}{v}\right) \\ &= \left(\rho v^2 \frac{1}{v} g_-\left(\frac{1}{v}\right), g_+\left(\frac{1}{v}\right)\right) \\ &= \left(\rho v g_-\left(v\right), g_+\left(v\right)\right) \\ &= Z\left(K\left(\rho,v\right)\right). \end{aligned}$$

The case for fast waves is similar. ∎

So when solving the Riemann problem we have three waves at our disposal: C_1-waves, C_2-waves, and K-waves. The K-waves always have zero speed, and the C-waves can have both positive and negative speed. The solution has to contain a K-wave to bring us from one part of the pipe to the other. This means that a priori the solution may consist of up to five different waves: $C_1C_2KC_1C_2$. However, the following lemma limits this to four.

LEMMA 2.2. *The wave configuration $\ldots C_2KC_1 \ldots$ is impossible.*

Proof. The smallest speed of the right edge of any C_2-wave with a right state (ρ_l, v_l) is $v_l + 1$, and the largest speed of the left edge of a C_1-wave with a left state (ρ_r, v_r) is $v_l - 1$. Therefore we have the inequalities

$$v_l + 1 \leq 0 \leq v_r - 1,$$

which means that $v_l \leq -1$ and $v_r \geq 1$. Since ρv is constant across K-waves, there can be no K-wave connecting (ρ_l, v_l) and (ρ_r, v_r). ∎

Consequently, we are left with the four possible wave configurations to solve the Riemann problem:

$$C_1KC_1C_2, \;\; C_1KC_2, \; C_1C_2KC_2, \text{ and } KC_1C_2.$$

When solving the Riemann problem, we use the strategy that for each fixed left state $L = (\rho_L, v_L)$ we construct waves with increasing speed until we reach the right state R. In this way we partition the (ρ, v) plane into regions such that for each $R = (\rho_R, v_R)$ in a region, the wave structure of the solution is constant. Hence, we may have a $C_1C_2KC_2$ region, a C_1KC_2 region, etc. States to the left of the K-wave are denoted by l and states to the right are denoted by r.

It turns out that there are two cases: $v_L > v_c^+$ and $v_L \leq v_c^+$.

Case 1. $v_L > v_c^+$. Consider first a right state R near L. The waves in a neighborhood of L all have positive speed, and hence we have to start with a K-wave. Let $r_1 = (\rho_{r_1}, v_{r_1}) = K(L)$. Then $v_{r_1} \in [1, v_L]$. We now construct wave curves $C_1(r_1)$ and $C_2(r_1)$. For $r_2 \in R_1(r_1)$ wave speeds are all positive, and we may continue with a fast wave $C_2(r_2)$ to reach a right state R, viz. $R \in C_2(r_2)$. If, however, $r_2 \in C_1(r_1)$, the shock speed is positive only down to the point $Z(r_1)$, which by Lemma 2.1 equals $K(Z(L))$. Thus for each $r_2 = (\rho_{r_2}, v_{r_2})$ with $v_{r_2} \in [1/v_{r_1}, v_{r_1}]$, we may continue with a fast wave to a right state $R \in C_2(r_2)$. In this way we fill the region denoted by KC_1C_2 above the curve $C_2(K(Z(L)))$. To prove uniqueness, consider a right state R in this region. The Riemann problem with left state $K(L)$ and right state R has a unique solution by standard techniques, and hence the only alternative would be to start from the right with a K-wave to the state $K^{-1}(R)$, which, however, is in the region with only waves with positive speed and hence is impossible.

Consider now a point l_1 on $S_1(L)$ between $Z(L)$ and $\widehat{L}$. L connects to l_1 with a slow shock with negative speed. For each such state l_1, we may continue with a K-wave to a state $r_1 = K(l_1)$. At the state $\widehat{L}$ the K-map ceases to be continuous, and the wave structure will be different. Fast rarefaction waves from r_2 will all have positive speed and may be used in the construction to reach a right state $R \in R_2(r_2)$. For fast shock waves emanating from r_1 with positive v_{r_1}, the shock speed remains positive (the Z map is not defined), and hence R may be any point on $S_2(r_1)$. If, however, r_1 is between $v = 0$ and $v = -v_c^-$, the shocks on $C_2(r_1)$ have positive speed down to $Z(r_1)$, and only states R above this point can be reached with this wave structure. Let c denote the part of $S_1(L)$ between $v = 0$ and $v = -v_c^-$, and let $\kappa = K(c)$, and finally $\zeta = Z(K(c))$. We then find that the solution reads C_1KC_2 in the region bounded from above by $C_2(K(Z(L)))$ and bounded from below by ζ and $C_2(K(\widehat{L}))$, the two latter curves starting from $Z(K(\widehat{L}))$. To prove uniqueness we first define $\tilde{c}$ as the part of $C_1(L)$ between $Z(L)$ and $\widehat{L}$, and subsequently $\tilde{\kappa} = Z(\tilde{c})$. First we have to prove that the curve $\tilde{\kappa}$ is transversal to C_2 curves starting from $\tilde{\kappa}$. This is the content of Lemma A.1 in the appendix. Furthermore, let R be above ζ. The only alternative to the given solution would be to connect the states $K^{-1}(R)$ and R instead of using a fast shock. As K^{-1} is monotone in the v variable, $K^{-1}(R)$ will be above $K^{-1}(\zeta) = Z(c)$. But then the state $K^{-1}(R)$ can be reached only with shocks with positive speed, making it impossible to end with a K-wave.

Consider now a right state R in the region denoted C_1C_2K, i.e., below the curves ζ and $v = -v_c^+$. Let $l_2 = K^{-1}(R)$. Then $v_{l_2} \leq -1$. The curves $C_2^-(l_2)$ and $S_1(L)$ intersect uniquely at a point l_1. If l_2 (yes, l_2) is below $S_1(L)$, the states l_1 and l_2 will connect using a rarefaction wave with negative speed, as $v_{l_2} \leq -1$. If, on the other hand, l_2 is below $S_1(L)$, we will use a fast shock wave to connect l_1 with l_2. As R is below ζ and the map K^{-1} is monotone in the v variable, $l_2 = K^{-1}(R)$ also will be below $K^{-1}(\zeta) = Z(c)$ using Lemma 2.1, and the shock will have negative speed as required. In this way we solve the Riemann problem in the region C_1C_2K.

Finally, the region $C_1C_2KC_2$ is bounded from above by the curve $C_2(K(\widehat{L}))$ and

by the line $v = -v_c^+$ from below, both curves emanating from $Z(K(\widehat{L}))$. Let R be in this region. Assume R is such that $v_R \geq -v_c^-$. Then we let $r = (\rho_r, -v_c^-)$ be the unique point such that $R \in R_2(r)$, and subsequently $l_2 = K^{-1}(r) = (\rho_{l_2}, -1)$. There is a unique state l_1 on $S_1(L)$ such that $l_2 \in R_2(l_1)$. Thus the solution consists of the following waves: The left state L is connected to state l_1 using a slow shock wave (with negative speed), followed by a fast rarefaction wave (with nonpositive speed) up to the state l_2 on the line $v = -1$. This state is connected with a K-wave to the state r. Finally, we use a fast wave to connect to R. If R has v_R less than $-v_c^-$, we start by using a fast shock with positive speed from $r = (\rho_r, -v_c^-)$ down to R, i.e., $R \in S_2(r)$. The remaining part of the solution is the same. For the uniqueness question, consider first the case with $v_R > -v_c^-$. Any waves connecting to $K^{-1}(R)$ would have positive speed, which would make it impossible to connect $K^{-1}(R)$ and R with a stationary K-wave. If v_R is between $-v_c^+$ and $-v_c^-$, we cannot apply $K^{-1}(R)$, and all fast waves have negative speed in that region, making it impossible to construct a different solution.

Case 2. $v_L \leq v_c^+$. If $v_L \leq v_c^+$, we can connect to the state $l_1 = \widetilde{L}$ on the intersection $C_1(L) \cap \{v = v_c^-\}$ with a wave of nonpositive speed. (If $v_L < v_c^-$ the wave is a rarefaction wave, and if $v_L > v_c^-$ it will be a shock wave.) l_1 can be connected to the state $r_1 = K(l_1)$ with a K-wave. The slow rarefaction wave starting from r_1 will have nonnegative speed, and hence we may use fast waves from any point $r_2 \in R_1(r_1)$ to reach a right state $R \in C_2(r_2)$. In this way we fill the region above the curve $C_2(K(\widetilde{L}))$ with a solution of the form $C_1KC_1C_2$. The remaining part of the construction is similar to that of Case 1.

The curves separating the various regions are illustrated in Figures 3 (Case 1) and 4 (Case 2). An illustration of solution curves in all cases is given in Figure 5: the left column for Case 1 and the right column for Case 2.

Hence we have proven the following result.

THEOREM 2.3. *Let $0 \leq k < 2$. Then the Riemann problem*

$$\begin{gathered} \rho_t + (\rho v)_x = 0, \\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x = -k\delta_0 \rho v, \\ \rho(x,0) = \begin{cases} \rho_L & \text{for } x < 0, \\ \rho_R & \text{for } x > 0, \end{cases} \qquad v(x,0) = \begin{cases} v_L & \text{for } x < 0, \\ v_R & \text{for } x > 0 \end{cases} \end{gathered} \tag{2.6}$$

has a unique solution in the class of combinations of Lax shocks, rarefaction waves, and K-waves for any left state (ρ_L, v_L) and right state (ρ_R, v_R) with positive densities ρ_L and ρ_R given by the above construction.

3. The Riemann problem for addition of momentum. The general structure of the argument is identical to that used in the case with k positive. However, the properties of the function $g_\pm$ and $f_\pm$ are different in the two cases (cf. Figure 2 and Figure 6), and a separate discussion is required. To keep the presentation short, we give the details only where they are different from those of the previous section. Define

$$v_c^\pm = g_\pm(1), \quad v_c^- \leq 1 \leq v_c^+.$$

The symmetry properties of $g_\pm$ imply that $g_\pm(-v_c^\pm) = -1$. Furthermore, we have that

$$\frac{dg_+(v)}{dv} \geq 0 \text{ for } v \in \langle -\infty, -v_c^+] \cup [1, \infty \rangle$$

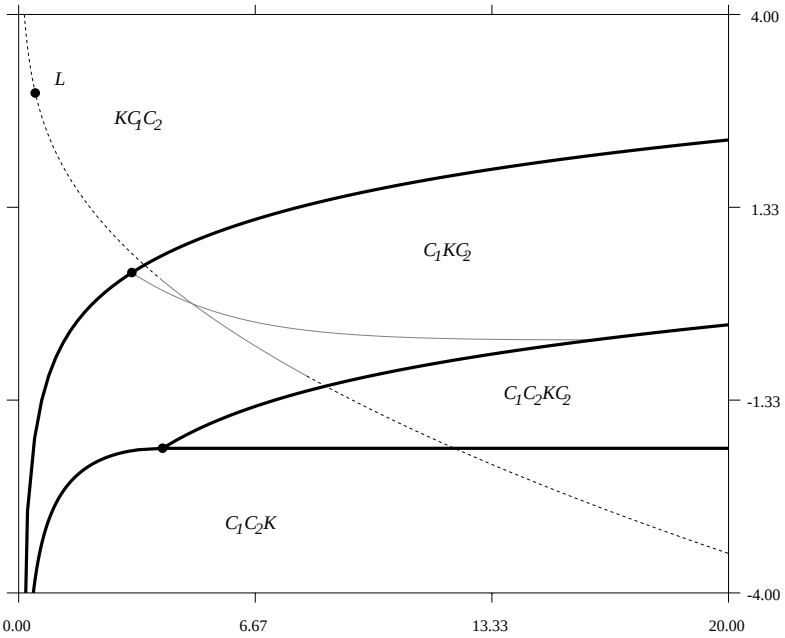


FIG. 3. *The solution to the Riemann problem where* $v_L \geq v_c^+$.

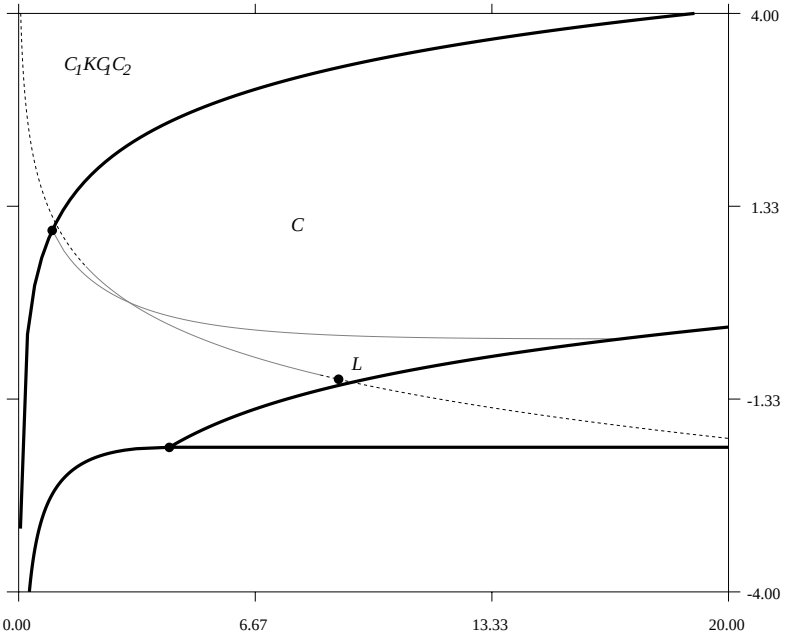


FIG. 4. *The solution to the Riemann problem where* $v_L \leq v_c^+$.

and

$$\frac{dg_-(v)}{dv} \geq 0 \text{ for } v \in [-v_c^+, 1],$$

and hence the K-wave reads

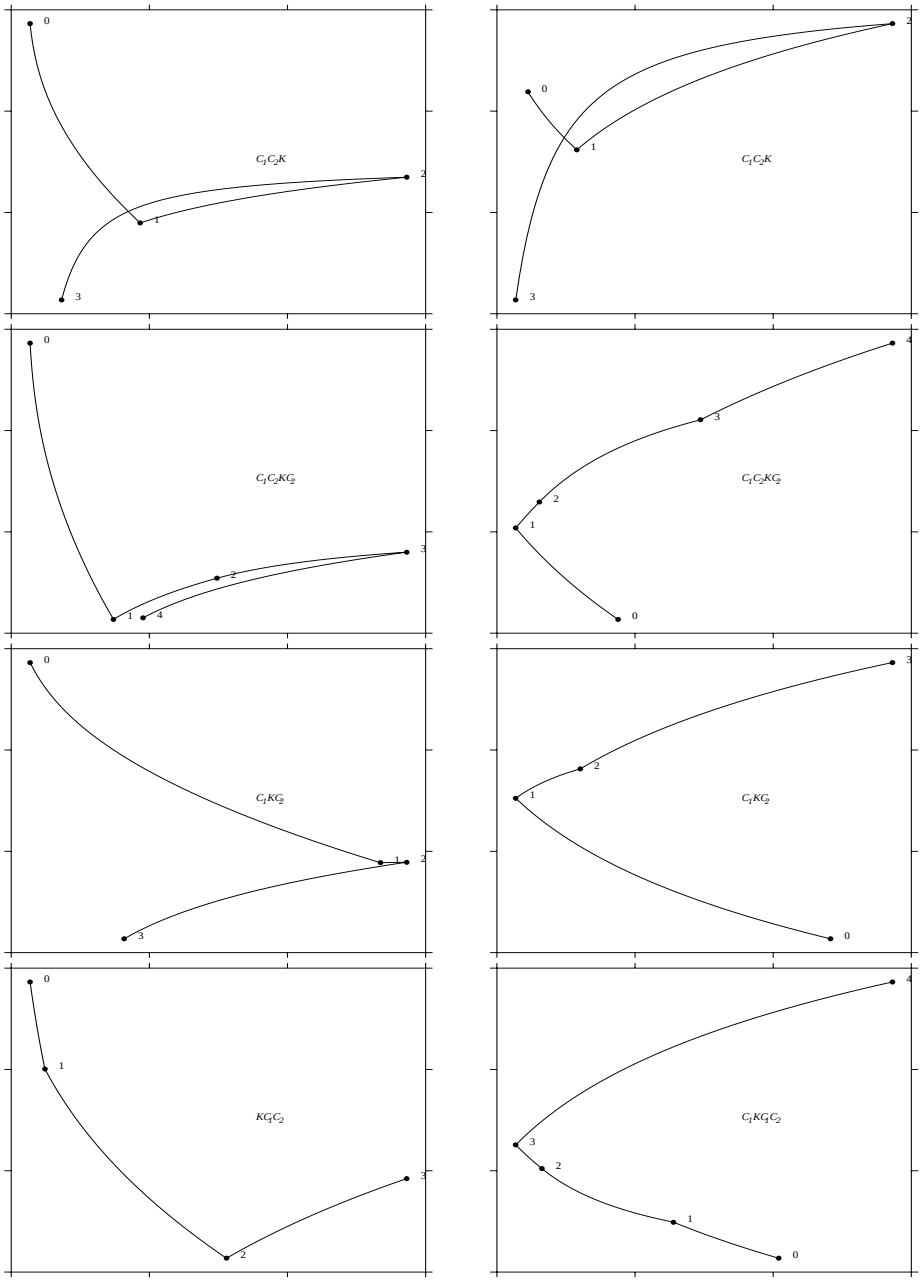


FIG. 5. *The solution in phase space in all subcases: $v_L \geq v_c^+$ right, $v_L \leq v_c^+$ left.*

$$K(\rho, v) = \begin{cases} (\rho f_+(v), g_+(v)) & \text{for } v \leq -v_c^+ \text{ and for } v \geq 1, \\ (\rho f_-(v), g_-(v)) & \text{for } v \in [-v_c^-, 1]. \end{cases} \tag{3.1}$$

The Z map (defined as before) and the K-map still commute, i.e., Lemma 2.1 remains valid. As designated points on the slow wave curve, it is convenient to define $\widehat{L}$ as the unique intersection of $C_1(L)$ and $\{v = -v_c^-\}$, and $\widetilde{L}$ as the unique intersection

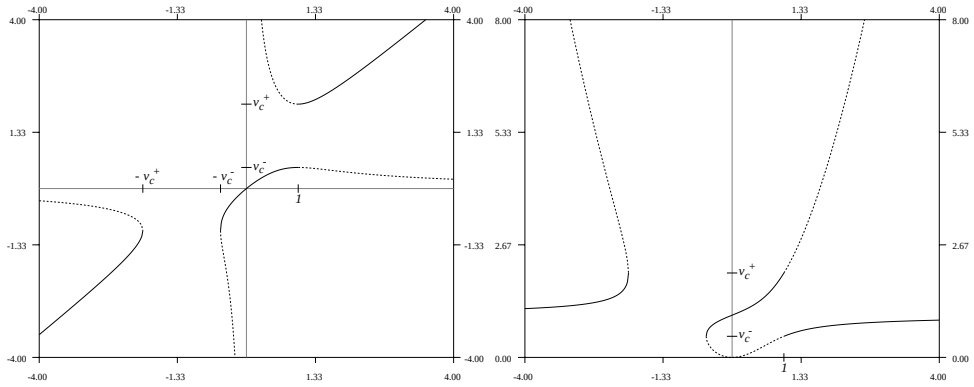


FIG. 6. *The functions* $g_\pm$ (*left*) *and* $f_\pm$ (*right*) *for* $k = -.5$.

between $C_1(L)$ and $\{v = 1\}$.

The general strategy for the construction of the solution of the Riemann problem is independent of the sign of k. Lemma 2.2 still reduces the number of possible wave combinations to four. As in the previous case, the analysis requires two distinct subcases; $v_L > 1$ and $v_L \leq 1$.

Case 1. $v_L > 1$. Waves in the neighborhood of L all have positive speed and we start with a K-wave to the state $r_1 = K(L)$. From the state r_1 we construct fast wave curves $C_1(r_1)$ and $C_2(r_1)$, which are all permissible except states on the slow shock below $Z(r_1) = K(Z(L))$. This results in a wave structure of the solution given by KC_1C_2 above the curve $C_2(K(Z(L)))$.

Now let l_1 be a state on $S_1(L)$ between $Z(L)$ and $\widehat{L}$. The left state L connects to l_1 with a shock with negative speed. We now employ the K-map to a state $r_1 = K(l_1)$. Fast rarefaction waves emanating from r_1 all have positive speed and can be used if the right state $R \in R_2(r_1)$. If v_{r_1} is positive, fast shocks on $S_2(r_1)$ will always have positive speed, while if $v_{r_1} \geq -1$ but negative, then the fast shocks will have positive speed only down to $Z(r_1)$. As before let c denote the part of $S_1(L)$ between $v = 0$ and $v = -v_c^-$, and $\kappa = K(c)$ and $\zeta = Z(\kappa)$. Then we obtain a solution with structure C_1KC_2 in the region bounded from above by $C_2(K(Z(L)))$ and below by ζ and $C_2(K(\widehat{L}))$ (starting from $Z(K(\widehat{L}))$). The question of uniqueness is more difficult in this case as we see that the curve $\tilde{\kappa} = K(\tilde{c})$, where $\tilde{c}$ is the part of $C_1(L)$ between $Z(L)$ and $\widehat{L}$, is in fact tangent to the curve $C_2(K(\widehat{L}))$. A proof that $\tilde{\kappa}$ is transversal to C_2 curves originating from $\tilde{\kappa}$ (except at $K(\widehat{L})$) is given in the appendix under the assumption that $k > 2(1 - \sqrt{2})$.

The construction of a solution with right state R in the region below ζ and $v = -1$ is similar to that in the case k positive.

Finally, let R be in the region bounded from above by $R_2(K(\widehat{L}))$ and from below by $v = -1$. Then there is a unique state r on $v = -1$ such that $R \in R_2(r)$, hence using only positive speeds. We have that r can be connected with K-wave from a state l_2 with $v_{l_2} = -v_c^+$. For each such state there is a unique state l_1 on $C_1(L)$ such that $l_2 \in C_2(l_1)$. This concludes the discussion of Case 1.

Case 2. $v_L \leq 1$. Because $v_L \leq 1$ we can use a one-rarefaction wave with non positive speed to reach the state $\widetilde{L}$ on the intersection of $C_1(L)$ and $v = 1$. When $v = 1$ we have two options for the map K, and denote $K^\pm(\rho, 1) = (\rho v_c^\mp, v_c^\pm)$. (For

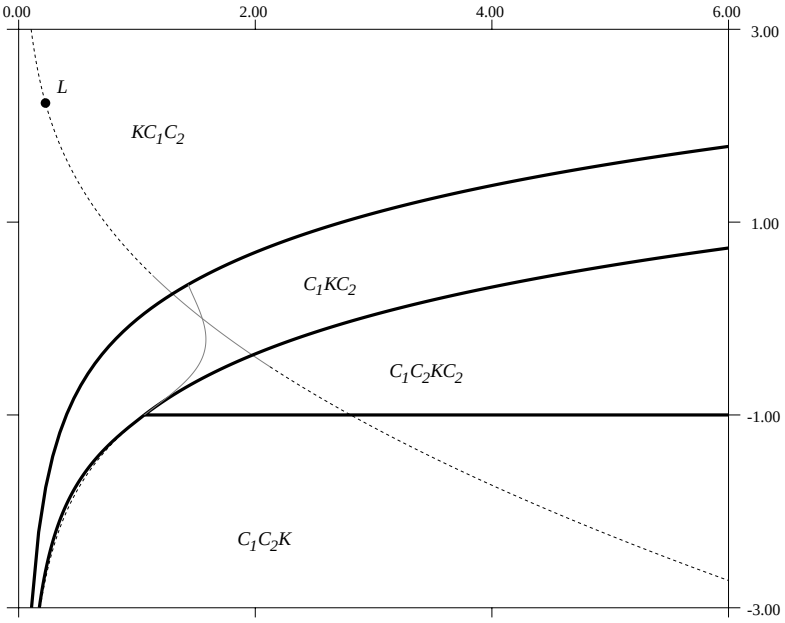


FIG. 7. *The solution of the Riemann problem for* $v_L > 1$.

the curve $\tilde{\kappa}$ we used the map K^-.) From the state $\widetilde{L}$ we use the K^+-map to the state $r_1 = K^+(\widetilde{L})$. We can now employ both slow and fast waves using $C_1(r_1)$ and $C_2(r_1)$. However, as in the case with k positive, we can use slow shocks only down to the point $Z(K^+(\widetilde{L}))$ which is equal to $K^-(\widetilde{L})$. Hence we obtain a solution structure of the form $C_1KC_1C_2$ in the region above $C_2(K^-(\widetilde{L}))$.

The remaining part of the construction equals that of Case 1.

We have proved the following theorem.

THEOREM 3.1. *Let* $0 \geq k > 2(1-\sqrt{2})$. *Then the Riemann problem*

$$\begin{gathered}\rho_t + (\rho v)_x = 0,\\ (\rho v)_t + \left(\rho v^2 + \rho\right)_x = -k\delta_0 \rho v,\\ \rho(x,0) = \begin{cases}\rho_L & \text{for } x<0,\\ \rho_R & \text{for } x>0,\end{cases} \qquad v(x,0) = \begin{cases}v_L & \text{for } x<0,\\ v_R & \text{for } x>0\end{cases}\end{gathered} \tag{3.2}$$

has a unique solution in the class of combinations of Lax shocks, rarefaction waves, and K-waves for any left state (ρ_L, v_L) *and right state* (ρ_R, v_R) *with positive densities* ρ_L *and* ρ_R *given by the above construction.*

See Figures 7 and 8 for an illustration of the various regions in Cases 1 and 2, respectively.

4. Conclusion. We have solved the Riemann problem for isothermal gas flow in a thin pipe through a sharp bend, a kink, modeled by a Dirac delta function located at the bend in the momentum equation. The equations then read

$$\rho_t + (\rho v)_x = 0, \quad (\rho v)_t + \left(\rho v^2 + \rho\right)_x = -k\delta_0 \rho v$$

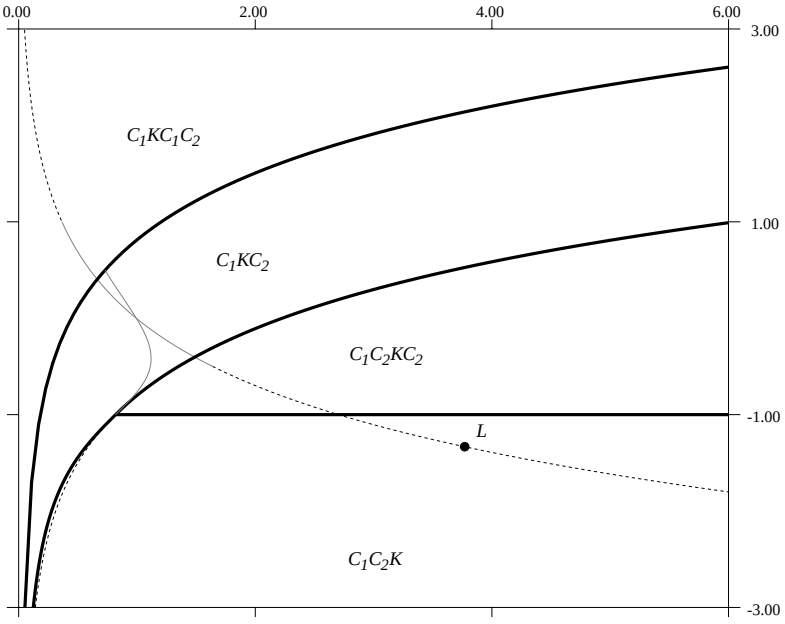


FIG. 8. *The solution of the Riemann problem for* $v_L < 1$.

with Riemann initial data. Here $k = \sqrt{2(1-\cos\theta)}$, where θ is the angle of the kink. By considering the case where k is positive, we model the case when momentum is added at an isolated point along the pipe. In both cases the Riemann problem has a unique solution.

Several important extensions are possible: First of all, the case of more general Cauchy initial data would be very interesting; i.e., both the source and the initial data are more general functions than those considered here. It is likely that the solution with Riemann initial data will play an important role, for instance, in building approximate solutions by Glimm's method or by front tracking.

Second, the extension to polytropic gas rather than isothermal gas would be quite interesting. In this case the algebraic manipulations are more complicated.

5. Appendix. In this appendix we prove that the curve $\tilde{\kappa}$ and C_2 curves starting from $\tilde{\kappa}$ are transversal. We first observe that for the C_2 curves originating on $\tilde{\kappa}$ we have that $d\tilde{\rho}/d\tilde{v} = \tilde{\rho}$, and hence it suffices to show that $d\tilde{\rho}/d\tilde{v}$ for the curve $\tilde{\kappa}$ is not equal to $\tilde{\rho}$. In the case k positive we show that indeed $d\tilde{\rho}/d\tilde{v}$ is negative, while in the case k is negative, the estimates have to be sharper as the curves are in fact tangent at the end point.

We start with the case when k is positive.

LEMMA A.1. *Assume that* $0 \le k \le 2$. *Let* $(\tilde{\rho}, \tilde{v})$ *be an arbitrary point on* $\tilde{\kappa}$. *Then*

$$\frac{d\tilde{\rho}}{d\tilde{v}} \le 0. \tag{A.1}$$

Proof. By the chain rule

$$\frac{d\tilde{\rho}}{d\tilde{v}} = \frac{\tilde{\rho}}{g_-'(v)}\left(\frac{\rho'}{\rho} + \frac{1}{v} - \frac{g_-'}{g_-}\right), \tag{A.2}$$

where we let the C_1 curve through (ρ_L, v_L) be parameterized by $(\rho(v), v)$. We have that $\rho'/\rho \le 0$, and the lemma will follow if

$$\frac{1}{v} - \frac{g_-'}{g_-} = \frac{g_- - g_-' v}{v g_- g_-'} \le 0. \tag{A.3}$$

The last inequality holds since g_-' satisfies the differential equation

$$g_-' = \frac{1 - v^2}{v\sqrt{\alpha^2 - 4v^2}} g_-, \tag{A.4}$$

which implies that $g_- - g_-' v \le 0$ as long as $k \le 4$. ∎

We now discuss the more complicated question of transversality of the $\tilde{\kappa}$ curve in the case with k negative.

LEMMA A.2. *Assume* $2(1-\sqrt{2}) < k < 0$. *Let* $(\tilde{\rho}, \tilde{v})$ *be an arbitrary point on* $\tilde{\kappa}$. *Then*

$$\frac{d\tilde{\rho}}{d\tilde{v}} < \tilde{\rho}, \tag{A.5}$$

except at the point $K(\widehat{L})$.

Proof. We parametrize the curve $\tilde{\kappa}$ using the parameter v running along the $C_1(L)$ curve, and hence write $\tilde{\rho} = \tilde{\rho}(\rho(v), v)$ and similarly for the other dependent variables.

We first compute

$$\begin{aligned}
\frac{d\tilde{\rho}}{d\tilde{v}} = \frac{\frac{d\tilde{\rho}}{dv}}{\frac{d\tilde{v}}{dv}} &= \frac{\frac{(\rho' v + \rho) g - \rho v g'}{g^2}}{g'} \\
&= \frac{1}{g'}\left(\frac{\rho' \rho v}{g\rho} + \frac{\rho v}{g v} - \frac{\rho v g'}{g g}\right) \\
&= \frac{\tilde{\rho}}{g'}\left(\frac{\rho'}{\rho} + \frac{1}{v} - \frac{g'}{g}\right).
\end{aligned} \tag{A.6}$$

What we aim to show is that

$$\frac{1}{v} + \frac{\rho'}{\rho} < \frac{g'}{g}(g+1), \tag{A.7}$$

except at the point $K(\widehat{L})$. From this it follows that $d\tilde{\rho}/d\tilde{v} < \tilde{\rho}$ away from the point $K(\widehat{L})$.

We have that

$$\frac{\rho'}{\rho} = \begin{cases} -1 & \text{for } \rho \le \rho_L, \\ -\frac{2\sqrt{\rho\rho_L}}{\rho + \rho_L} & \text{for } \rho > \rho_L. \end{cases}$$

We first show that $\tilde{\kappa}$ is never tangent to the C_2 curves for $v > 0$. Now we need the following lemma, the proof of which comes after this proof.

LEMMA A.3. *Let* $(\rho, v) = (\rho(v), v) \in C_1(L)$. *Then*

$$\frac{\rho'}{\rho} < \frac{-2v}{v^2+1} \tag{A.8}$$

for $0 < v < 1$.

With (A.7) in mind we introduce the reparametrization β by

$$\beta(u) = u - k + \frac{1}{u},$$
$$u = \frac{1}{2}\left(\gamma - \sqrt{\gamma^2 - 4}\right),$$

where $\gamma = \beta + k \geq 2$. The parameter β is in the interval $[2-k, \infty\rangle$. We have that $d\beta/dv = 1 - 1/v^2 < 0$, in this interval. Therefore, (A.7) will follow if

$$V(v) := \frac{\frac{1}{v} - \frac{2v}{v^2+1}}{1 - \frac{1}{v^2}} = \frac{-v}{v^2+1} = -\frac{1}{\gamma}$$

is greater than

$$H(v) := \frac{\frac{dg}{d\beta}}{g}(g+1) = -\frac{1}{2}\left(\sqrt{\frac{\beta+2}{\beta-2}} - 1\right).$$

Solving the equation $H = V$ for $\gamma = \tilde{\gamma}(k)$, we obtain

$$\tilde{\gamma}(k) = \frac{2+k}{k+1}. \tag{A.9}$$

Since $\gamma \geq 2$, and $\tilde{\gamma}(k) < 0$ for $k > -1$, we have that $H < V$ for all $k > -1$. This concludes the case $v > 0$.

For the (harder) case where $v < 0$, we use the crude estimate $\rho'/\rho < 0$ and wish to prove

$$\frac{1}{v} < \frac{g'}{g}(g+1). \tag{A.10}$$

Now we use the reparametrization

$$\beta = -v + k - \frac{1}{v},$$
$$v = -\frac{1}{2}\left(\gamma - \sqrt{\gamma^2 - 4}\right),$$

where $\gamma = \beta - k$. Now β is in the interval $[2, \infty\rangle$. We have that $d\beta/dv = 1/v^2 - 1 > 0$; consequently, (A.10) will follow if we can show that

$$V(v) := \frac{\frac{1}{v}}{\frac{1}{v^2} - 1} = \frac{v}{1 - v^2} = \frac{-1}{\sqrt{\gamma^2 - 4}}$$

is less than

$$H(v) := \frac{\frac{dg}{d\beta}}{g}(g+1) = \frac{1}{2}\left(\sqrt{\frac{\beta-2}{\beta+2}} - 1\right).$$

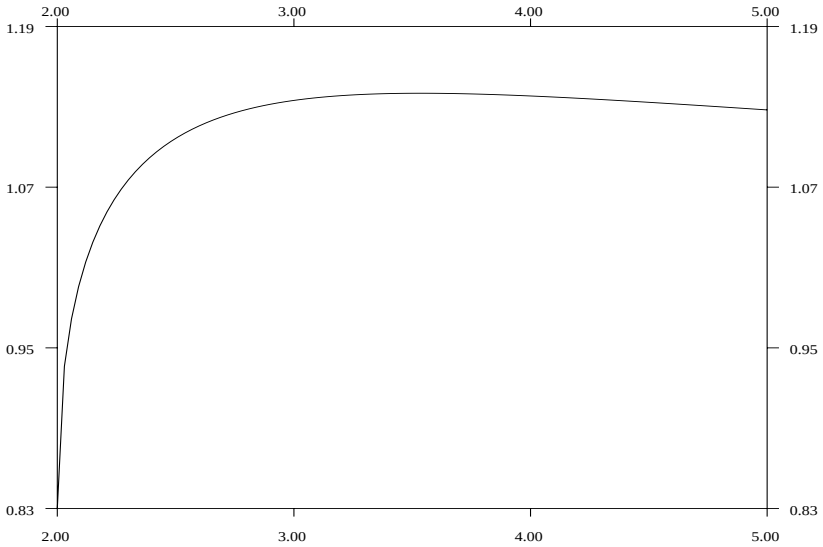


FIG. 9. *The function* $\tilde{k}(\beta)$.

In this case we can solve the equation $V = H$ for $k = \tilde{k}(\beta)$, obtaining

$$\tilde{k}(\beta) = -\beta + \sqrt{\frac{1}{2}(\beta+2)^2\left(\sqrt{\frac{\beta-2}{\beta+2}}+1\right) - (\beta-2)}\,.$$

For β in the interval $[2,\infty\rangle$, $\tilde{k}(\beta) \geq \tilde{k}(2) = 2(1-\sqrt{2})$; see Figure 9 below. Hence, for $k > 2(1-\sqrt{2})$, $V(v) < H(v)$. □

Proof of Lemma A.3. If (ρ, v) is on the rarefaction part of the C_1 curve, then $\rho'/\rho = -1$, and the lemma certainly holds. Assume therefore that (ρ, v) is on the shock part of the C_1 curve. Then it is below $Z(L) = (\rho_L v_L^2, 1/v_L)$, and hence $v \leq 1/v_L$. In this case we have that

$$\frac{\rho'}{\rho} = \frac{-2\sigma}{\sigma^2+1}, \tag{A.11}$$

where $\sigma(v) = \sqrt{\rho_0/\rho(v)}$. We see that $\sigma' \leq 1/2$, and $\sigma(1/v_L) = 1/v_L$, showing that $\sigma(v) > v$ for $v \leq 1/v_L$. The function $\phi(t) = -2t/(t^2+1)$ is monotonically decreasing for $|t| \leq 1$, and hence we infer that

$$\frac{\rho'}{\rho} < \frac{-2v}{v^2+1}. \qquad \square$$

REFERENCES

[1] 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.

[2] P. GLAISTER, *Flux difference splitting for hyperbolic systems of conservation laws with source terms*, Comput. Math. Appl., 26 (1993), pp. 79–96.

[3] J. M. GREENBERG, A. Y. LEROUX, R. BARAILLE, AND A. NOUSSAIR, *Analyse et approximation de lois de conservation avec terme source*, C. R. Acad. Sci. Paris Sér. I Math., 321 (1995), pp. 1073–1076.

[4] T. P. Liu, *Quasilinear hyperbolic systems*, Comm. Math. Phys., 68 (1979), pp. 141–172.
[5] T. P. Liu, *Non-linear stability and instability of transonic gas flow through a nozzle*, Comm. Math. Phys., 83 (1982), pp. 243–260.
[6] G. Crasta and B. Piccoli, *Viscosity solutions and uniqueness for systems of inhomogeneuous balance laws*, Discrete Contin. Dynam. Systems, 3 (1997), pp. 477–502.
[7] C. Z. Li and T. P. Liu, *Asymptotic states for hyperbolic conservation laws with a moving source*, Adv. in Appl. Math., 4 (1983), pp. 353–379.
[8] T. P. Liu, *Nonlinear resonance for quasilinear hyperbolic equation*, J. Math. Phys., 28 (1987), pp. 2593–2602.
[9] D. Marchesin and P. J. Paes-Leme, *A Riemann problem in gas dynamics with bifurcation*, Comput. Math. Appl., 12 (1986), pp. 433–455.
[10] T. Gimse and N. H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.
[11] G.-Q. Chen and J. Glimm, *Global solutions to the compressible Euler equations with geometrical structure*, Comm. Math. Phys., 180 (1996), pp. 153–193.
[12] E. Isaacson and B. Temple, *Nonlinear resonance in systems of conservation laws*, SIAM J. Appl. Math., 52 (1992), pp. 1260–1278.
[13] E. Isaacson, *Global Solution of a Riemann Problem for a Non-strictly Hyperbolic System of Conservation Laws Arising in Enhanced Oil Recovery*, preprint, Rockefeller University, New York, 1981.
[14] J. Smoller, *Shock Waves and Reaction–Diffusion Equations*, Springer, New York, 1994.
[15] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 1992.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 516–526

RECOVERING ASYMPTOTICS OF COULOMB-LIKE POTENTIALS FROM FIXED ENERGY SCATTERING DATA*

M. S. JOSHI†

Abstract. Any compact smooth manifold, X, with boundary admits a Riemannian metric of the form $x^{-4}dx^2 + x^{-2}h'$ near the boundary with x a boundary defining function and h' restricting to a metric on the boundary. Melrose [*Spectral and scattering theory for the Laplacian on asymptotically Euclidean spaces*, in Spectral and Scattering Theory, M. Ikawa, ed., Marcel Dekker, New York, 1994] has associated a scattering matrix to such metrics and potentials in $xC^\infty(X)$. It is shown for potentials of the form $Ax + O(x^2)$ that this scattering matrix is a Fourier integral operator and that the asymptotics of such potentials can be recovered from the scattering matrix for various manifolds including Euclidean space.

Key words. scattering theory, Coulomb-like potentials, Lagrangian

AMS subject classification. 58G15

PII. S003614109732763X

1. Introduction. Our purpose in this article is to examine the microlocal properties of the scattering matrix at fixed energy associated to a class of potentials, which are similar to the Coulomb potential and show that the total symbols of these operators determine the asymptotics of these potentials. We shall work in the general context of a manifold with boundary equipped with a scattering metric which contains the important special case of Euclidean space. Our approach will be to extend the arguments of [3] and [5] to cover this more general case. In particular, we show that the scattering matrix at fixed energy is still a Fourier integral operator associated with geodesic flow at time π of order 0 but now with a classical symbol of imaginary order.

Recall, from [4], that a scattering metric on the manifold with boundary $(X, \partial X)$ is a metric which can be written in the form

$$g = \frac{dx^2}{x^4} + \frac{h}{x^2} \tag{1.1}$$

for some boundary defining function x, with h a symmetric tensor restricting to a positive definite form on $T(\partial X)$. (The choice of x is actually fixed up to $O(x^2)$ by g.) A long range potential is then a potential in the class $xC^\infty(X)$ and we shall define a Coulomb-like potential to be a long range potential of the form $Ax + O(x^2)$ for some $A \in \mathbb{R}$. Let Δ denote the Laplacian associated with g; then for each $f \in C^\infty(\partial X)$ and $\lambda \neq 0$ there is a unique eigenfunction u such that

$$(\Delta + V - \lambda^2)u = 0$$

of the form

$$e^{\frac{i\lambda}{x}} x^{\frac{n-1}{2} + i\alpha} f' + e^{\frac{-i\lambda}{x}} x^{\frac{n-1}{2} - i\alpha} f'',$$

*Received by the editors September 22, 1997; accepted for publication (in revised form) April 29, 1998; published electronically March 19, 1999.
http://www.siam.org/journals/sima/30-3/32763.html
†Department of Pure Mathematics and Mathematical Statistics, University of Cambridge, 16 Mill Lane, Cambridge CB2 1SB, England, UK (joshi@dpmms.cam.ac.uk).

where $\alpha = \frac{A}{2\lambda}$ with $f', f'' \in C^\infty(X)$, and such that the restriction of f' to the boundary is f. The scattering matrix associated with g and V is then defined to be the map

$$S(\lambda) : f \mapsto f'_{|\partial X}.$$

In the special case where $A = 0$, it is shown by Melrose and Zworski [5] that $S(\lambda)$ is a classical Fourier integral operator of order 0. The question of how to recover the asymptotics of V from the total symbol of $S(\lambda)$ in that case is addressed by Barreto and Joshi [3]. There, it is shown that given two potentials equal up to order k, the difference of the scattering matrices is order $1 - k$. The principal symbol of the difference of the associated scattering matrices was linearly determined by and determines weighted integrals of the lead term of the difference of the potentials along geodesics. It was also shown that a function could be recovered from these weighted integrals and thus that the asymptotics of a potential could be recovered from the scattering matrix.

Here we show that the analogous results hold when A is nonzero.

THEOREM 1.1. *Let $(X, \partial X, g)$ be manifold with boundary, with scattering metric g, and boundary defining function x. Let V be a Coulomb-like potential on $(X, \partial X)$ with lead term Ax; then the scattering matrix at energy λ associated with g and V is an elliptic, classical Fourier integral operator of order $-iA/\lambda$.*

Note that a classical Fourier integral operator of order $-iA/\lambda$ is of course a Fourier integral operator of order 0.

THEOREM 1.2. *Let $(X, \partial X, g, x)$ be as above. Let V_1, V_2 be Coulomb-like potentials on X such that*

$$V_1 - V_2 = x^k W + O(x^{k+1})$$

for $W \in C^\infty(\partial X)$; then if $S_j(\lambda)$ is the scattering matrix associated with V_j we have that $S_1(\lambda) - S_2(\lambda)$ is a classical Fourier integral operator of order $-iA/\lambda + 1 - k$ and the principal symbol of $S_1(\lambda) - S_2(\lambda)$ determines and is determined by

$$\int_0^\pi \sin(s)^{k-1} W(\gamma(s)) ds$$

for all geodesics γ of length π on ∂X.

The boundary ∂X has a natural Riemannian structure so the X-ray transform of a function on it is well defined. The x-ray transform is the operator on the space of functions on ∂X to the space of functions on the space of closed geodesics obtained by integrating along the geodesics. We recall that a manifold has an injectivity radius of at least R if the exponentiation map from the tangent space to the manifold is injective for vectors of size up to R. Then the same argument as in [3] is shown below.

THEOREM 1.3. *If $(X, \partial X)$ is such that the X-ray tranform on ∂X is injective, or the injectivity radius of ∂X is greater than π, or ∂X is a sphere not of radius $\frac{1}{k+1}$, $k > 0$, then if the Coulomb-like potentials V_1, V_2 have scattering matrices at some energy which are equal up to smoothing then $V_1 - V_2 = O(x^\infty)$.*

In particular, scattering data determine the asymptotics on Euclidean space as it is a special case, where the boundary is a sphere which is of radius 1.

The problem of recovering the short range part of a long range potential has been studied by Isozaki and Kitada [2] in the 2-body case and Enss and Weder [1] in the n-body case. They proved that the short range part is uniquely determined by the high energy limit of the scattering matrix.

It is known that one cannot recover Schwartz potentials, and examples of transparent potentials at fixed energy have been constructed by Regge [7], Newton [6], Sabatier [8], and others. Thus one would not expect to do better than the asymptotics in the long range case.

Vasy [9] has informed us of an alternative proof of Theorem 1.1, which avoids the necessity of using a symbol calculus and therefore does not prove Theorems 1.2 and 1.3.

2. Review of definition of scattering matrix. In this section, we review some work of Melrose [4] on how to define the scattering matrix of a Coulomb-like potential on a general manifold with boundary equipped with a scattering metric. Suppose that a boundary defining function x has been chosen and that $V = Ax + O(x^2)$. We work in a fixed product decomposition near the boundary. The Laplacian is then of the form

$$(x^2 D_x)^2 + ix(n-1)x^2 D_x + x^2 \Delta_{\partial X} + R,$$

where R is lower order at the boundary. Then

$$(\Delta + V - \lambda^2)(x^p e^{i\lambda/x} b(y)) = i\lambda(2p - n + 1 - iA/\lambda) x^{p+1} e^{i\lambda/x} b + O(x^{\Re p+1}).$$

Thus if we take $p = \frac{n-1}{2} + i\alpha$ with $\alpha = A/2\lambda$, the lead term vanishes and we can iteratively solve away the error terms and, applying the Borel lemma, obtain $f(x,y) \in C^\infty(X)$ such that $(\Delta + V - \lambda^2)(e^{i\lambda/x} x^p f(x,y))$ vanishes to infinite order at $x = 0$ and $f(0,y) = b(y)$. With this modification, the arguments in [4] are resolved and there is a unique eigenfunction of the form

$$e^{\frac{i\lambda}{x}} x^{\frac{n-1}{2} + i\alpha} f + e^{-\frac{i\lambda}{x}} x^{\frac{n-1}{2} - i\alpha} f'$$

with $f' \in C^\infty(X)$. The scattering matrix at energy λ is then the map

$$S(\lambda) : b \mapsto f'_{|\partial X}.$$

It is a unitary operator.

3. Review of Legendrian distributions. In this section, we review and rephrase the material we need from [4] and [5]. In this section, X is a compact manifold with boundary ∂X and g is a scattering metric on X with x a boundary defining function for ∂X such that g takes the form (1.1). Our account is necessarily brief and we refer the reader to [4] and [5] for more details.

There is a natural bundle over X called the scattering cotangent bundle, which is denoted ${}^{sc}T^*(X)$. This is the dual to the bundle of smooth, vector fields of bounded length with respect to some (and hence all) scattering metrics on X. The restriction of ${}^{sc}T^*(X)$ to ∂X is denoted ${}^{sc}T^*(X)_{|\partial X}$ and carries a natural contact structure. If y are local coordinates on ∂X and μ are the corresponding dual coordinates, then (y, μ, τ) form local coordinates on ${}^{sc}T^*(X)_{|\partial X}$, where τ is the coefficient of $\frac{dx}{x^2}$. We assume that a product decomposition close to the boundary has been chosen $X = \partial X \times \mathbb{R}_+$. If y are coordinates on ∂X, we then have coordinates $\partial X \times \mathbb{R}_+$ on X close to the boundary.

We omit the definition of scattering pseudodifferential operators as we need only consider differential operators; however, we note that a differential operator $P(x, y, xD_y, x^2 D_x)$ will be in $\Psi_{sc}^{m,k}(X, {}^{sc}\Omega^{1/2})$ if it is of order m and that the total

symbol as an operator in xD_y, x^2D_x vanishes to kth order at the boundary. The operator P then has a well-defined symbol at the boundary:

$$j(P) = x^k p_k + x^{k+1} p_{k+1},\ p_k, p_{k+1} \in C^\infty(\mathbb{R} \times T^*\partial X).$$

DEFINITION 3.1. *An intersecting pair with conic points is a subset, $\widetilde{W}$, of ${}^{sc}T^*(X)_{|\partial X}$, which is a union of the closure of a smooth Legendrian submanifold, W, and $W^\#$, a finite union of global sections of the form $W^\#(\lambda_j) = \{(y, 0, \lambda_j)\}$, containing $\overline{W} \setminus W$. We also require $\overline{W}$ to have an at most conic singularity at $\mu = 0$; that is, it is smooth if polar coordinates are introduced along $\overline{W} \setminus W$.*

The process of introducing polar coordinates along $\overline{W} \setminus W$ can be given an invariant meaning and is then called blow-up. We denote the blown-up manifold $\widehat{W}$.

The metric g induces a metric h on the nearby boundary. It is of the form

$$\tau^2 + h'(y, \mu) + xg'$$

as a function on ${}^{sc}T^*X$; we obtain h' from h via the isomorphism

$$\mu.\frac{dy}{x} \longmapsto \mu.dy.$$

EXAMPLE 3.1. *For each $y' \in \partial X$ and $0 \neq \lambda \in \mathbb{R}$, let $G_{y'}(\lambda)$ be equal to the set of (τ, y, μ), such that $\tau^2 + |\mu|^2 = \lambda^2, \mu \neq 0$, and, putting $\mu = |\mu|\hat{\mu}$,*

$$\begin{aligned} \tau &= |\lambda| \cos(s), \\ |\mu| &= |\lambda| \sin(s), \\ (y, \hat{\mu}) &= \exp(sH_{\frac{1}{2}h})(y', \hat{\mu}'), \end{aligned} \tag{3.1}$$

where $s \in (0, \pi)$, $(y', \hat{\mu}') \in T^*\partial X$, and $h(y', \hat{\mu}') = 1$. Then $G_{y'}(\lambda) \cup \{(\lambda, y, 0)\}$ is an intersecting pair with conic points. We denote this pair $\widetilde{G}(\lambda)$. This is the pair we are interested in this paper. The set $G^\sharp(\lambda) = \{(-\lambda, y, 0, y', 0)\}$ is also important in our construction. Note that $G^\sharp(\lambda)$ is the initial or incoming surface and that $G^\sharp(-\lambda)$ is the outgoing surface. Note that in the coordinates defined by (3.1)*, $G^\sharp(\lambda)$ is $s = \pi$ and $G^\sharp(-\lambda)$ is $s = 0$, when λ is positive.*

Associated with these intersecting pairs at each conic point is a unique homogeneous Lagrangian submanifold $\Lambda(\widetilde{W}, \lambda_i)$ of $T^*(\partial X)$. For the pair we are interested in, $\widetilde{G}(\lambda)$ this is precisely the relation of being π apart along a lifted geodesic (see Proposition 4 of [5]). For simplicity, we shall henceforth take λ to be positive, the λ negative case is similar (or could be deduced from the positive case).

Melrose and Zworski [5] associated any such intersecting pair with a class of smooth functions whose asymptotics on approach to the boundary are determined by symbols on the Legendrians. A symbol bundle over the smooth Legendrian $W(\lambda)$ in pair $\widetilde{W}$ can be defined and is denoted $\hat{E}^{m,p}$. The sections of this bundle are of the form

$$aS^{p-m}|dx|^{m-n/4}$$

with a a smooth section of $C^\infty(\hat{W}; \Omega_b^{\frac{1}{2}} \otimes M_{\hat{H}})$, S a defining function of the boundary of W, M the Maslov bundle, and $\Omega_b^{\frac{1}{2}}$ the $b-$half density bundle. For G above, one could take $S = \sin s$. Melrose and Zworski remove this singularity at the endpoints by rescaling, but for us it will be easier not to do so.

PROPOSITION 3.1. *If $\widetilde{W}(\lambda)$ is an intersecting pair with conic points then there is a class of smooth half-densities on X^o, denoted $I^{m,p}_{sc}(X,\widetilde{W})$, such that $\bigcap_{m,p} I^{m,p}_{sc}(X,\widetilde{W})$ is equal to the class of half-densities vanishing to infinite order at the boundary. There exists a symbol map*

$$\hat{\sigma}_{sc,m,p} : I^{m,p}_{sc}(X,\widetilde{W},{}^{sc}\Omega^{1/2}) \to C^\infty(\hat{W};\hat{E}^{m,p})$$

which gives a short exact sequence

$$0 \to I^{m+1,p}_{sc}(X,\widetilde{W},{}^{sc}\Omega^{1/2}) \to I^{m,p}_{sc}(X,\widetilde{W},{}^{sc}\Omega^{1/2}) \to C^\infty(\hat{W};\hat{E}^{m,p}) \to 0.$$

This is Proposition 12 from [5]. An important related fact we need to know is, how do Legendrian distributions map under scattering pseudodifferential operators? We recall Proposition 13 from [5].

PROPOSITION 3.2. *Suppose $P \in \Psi^{l,k}_{sc}(X,{}^{sc}\Omega^{1/2})$ has symbol $x^k p_k + x^{k+1}p_{k+1}$ with respect to a product decomposition of X near ∂X, and suppose that*

$$W \subset {}^{sc}T^*_{\partial X}(X)$$

is a smooth Legendre submanifold. Then for any $m \in \mathbb{R}$,

$$P : I^m_{sc}(X,W;{}^{sc}\Omega^{1/2}) \to I^{m+k}_{sc}(X,W;{}^{sc}\Omega^{1/2}), \tag{3.2}$$

$$\sigma_{sc,m+k}(Pu) = (p_{k|G})\sigma_{sc,m}(u) \otimes |dx|^k. \tag{3.3}$$

Furthermore, if p_k vanishes identically on W, then

$$P : I^m_{sc}(X,W;{}^{sc}\Omega^{1/2}) \to I^{m+k+1}_{sc}(X,W;{}^{sc}\Omega^{1/2})$$

and

$$\sigma_{sc,m+k+1}(Pu)$$
$$= \left(\frac{1}{i}\left(L_V + \left(\frac{1}{2}(k+1) + m - \frac{n}{4}\right)\frac{\partial p_k}{\partial \tau}\right) + p_{k+1|W}\right) a \otimes |dx|^{m+k+1-\frac{n}{4}},$$

where $\sigma_{sc,m}(u) = a \otimes |dx|^{m-\frac{n}{4}}$ and V is the rescaled Hamiltonian vector field associated with p_k.

We omit the definition of the rescaled Hamiltonian vector field but recall that for the pair G we are studying, it is equal to

$$2\lambda \sin s \frac{\partial}{\partial s}$$

in the semiglobal coordinates given by (3.1).

We also need two push-forward theorems, Propositions 16 and 17 from [5]. They relate the singularities of the scattering matrix to the asymptotics in small x of the Poisson operator. Given a product decomposition near the boundary, there is a natural pairing

$$B : C^{-\infty}(X,{}^{sc}\Omega^{1/2}) \times C^\infty(\partial X;{}^{sc}\Omega^{1/2}) \to C^{-\infty}([0,\epsilon),{}^{sc}\Omega^{1/2}), \tag{3.4}$$

$$B(u,f) = x^{\frac{n-1}{2}} \int_{\partial X} u(x,y)f(y). \tag{3.5}$$

PROPOSITION 3.3. *For any intersecting pair of Legendre submanifolds with conic points W, the partial pairing* (3.5) *gives a map*

$$B : I^{m,p}_{sc}(X, \widetilde{W}; {}^{sc}\Omega^{1/2}) \times C^\infty(\partial X; {}^{sc}\Omega^{1/2}) \mapsto \sum_j I^{p+\frac{n-1}{4}}([0,\epsilon), W'(\bar\tau_j; {}^{sc}\Omega^{1/2})),$$

where the $W'(\bar\tau_j) = \{(0, -\tau_j dx/x^2)\}$ are the Legendre submanifolds corresponding to the components of $W^\#$ and

$$B(u,f) = \sum_j e^{-i\bar\tau/x} x^{p+n/4} Q^0_{\bar\tau_j}(u,f) \left|\frac{dx}{x^2}\right|^{\frac12} + O(x^{p+n/4+1})$$

with

$$Q^0_{\bar\tau}(u) \in I^{p-m-\frac{n-1}{4}}_{phg}(\partial X, \Lambda(\widetilde{W}, \bar\tau)),$$

and the principal symbol of $Q^0_{\bar\tau}(u)$ determines and is determined by the lead singularity of the principal symbol of u on W on approach to $W'(\bar\tau_j)$.

When the Legendrian distribution is actually associated with a smooth Legendrian submanifold the push-forward becomes much simpler, and this simplifies the construction of the Poisson operator.

PROPOSITION 3.4. *If G is a smooth Legendre variety and $u \in I^m_{sc}(X, G'^{sc}\Omega^{1/2})$ near $\tau = \bar\tau$, then the distribution $Q^0_{\bar\tau}$ is a Dirac delta distribution.*

4. Construction of the Poisson operator. In this section, we apply the calculus reviewed in section 3 to construct the Poisson operator and prove Theorem 1.1. We shall refer heavily to [5, section 15], as our construction is a modification of the one there.

We assume a product decomposition of X close to the boundary has been chosen and is fixed throughout this section. We then have as in [5] that Δ, the intrinsic Laplacian acting on scattering half-densities on $X \times \partial X$, induces an operator

$$\Delta_X \in \mathrm{Diff}^2_{\mathrm{sc}}(X \times \partial X, {}^{sc}\Omega^{1/2}(X \times \partial X))$$

by

$$\Delta_X \left(u \left|\frac{dx}{x^2}\right|^{\frac12} \left|\frac{dy}{x^{n-1}}\right|^{\frac12} \left|\frac{dy'}{x^{n-1}}\right|^{\frac12} \right) = \Delta \left(u(.,y) \left|\frac{dx}{x^2}\right|^{\frac12} \left|\frac{dy}{x^{n-1}}\right|^{\frac12} \right) \left|\frac{dy'}{x^{n-1}}\right|^{\frac12},$$

where $(x, y; y')$ is a point in $X \times \partial X$. Throughout this section V will be a Coulomb-like potential, that is,

$$V = Ax + x^2 f, f \in C^\infty(X),$$

and α will be $\frac{A}{2\lambda}$.

We recall the following from [3].

LEMMA 4.1. *The symbol at the boundary of Δ_X is $p = p_0 + xp_1$ with p_1 equal to $-i(n-1)\tau + c$, where c is the derivative of the metric at the boundary and of the form $c(y,\mu,\tau) = \tau f(y,\mu) + g(y,\mu)$ with f linear in μ and g quadratic in μ.*

We also note the following lemma, which is important in our construction to show that the transport equations are solvable.

LEMMA 4.2. *If* $L \in I^{m,-\frac{1}{4}+i\alpha}(X \times \partial X, \tilde{G}(\lambda), {}^{sc}\Omega^{1/2})$ *is such that*

$$(\Delta_X - \lambda^2 + V)L \in I^{m+2,\frac{3}{4}+i\alpha}(X \times \partial X, \tilde{G}(\lambda), {}^{sc}\Omega^{1/2}),$$

then

$$(\Delta_X - \lambda^2 + V)L \in I^{m+2,\frac{7}{4}+i\alpha}(X \times \partial X, \tilde{G}(\lambda), {}^{sc}\Omega^{1/2}).$$

This is a modification of Lemma 15 from [5], and in fact, the $\alpha = 0$ case is essential to the construction there also.

Proof. The proof is no different from that of Lemma 15 in [5], the only difference being that in (15.17), there appears an extra term α, which is canceled by the lead term of the potential. □

PROPOSITION 4.1. *For any* $0 \neq \lambda \in \mathbb{R}$ *there exists*

$$K \in I^{m,p_1,p_2}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2}),$$

such that

$$(\Delta_X - \lambda^2 + V)K \in \mathcal{C}^\infty(X \times \partial X; {}^{sc}\Omega^{1/2})$$

and

$$Q^0_\lambda(K) = \mathrm{Id}$$

with

$$m = -\frac{2n-1}{4} + i\alpha,\ p_1 = -\frac{1}{4} + i\alpha,\ p_2 = -\frac{1}{4} - i\alpha,$$

and the principal symbol of K *on* G *is*

$$C \sin(s)^{\frac{n-1}{2} - i\alpha} \tan(s/2)^{-i\alpha} \frac{|ds|^{\frac{1}{2}} |dy|^{\frac{1}{2}} |d\hat{\mu}|^{\frac{1}{2}}}{(\sin s)^{\frac{1}{2}}} |dx|^{m - \frac{2n-1}{4}},$$

where $C(y, \hat{\mu})$ *is a nonzero smooth function.*

Note that in contrast to [5], we allow different orders on $G^\sharp(\lambda)$ and $G^\sharp(-\lambda)$.

Proof. As in [5], we first construct $K^b \in I^{m,p_2}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2})$ such that

$$(\Delta_X - \lambda^2 + V)K^b \in I^{\frac{3}{4} - i\alpha}_{sc}(X \times \partial X, G^\sharp(-\lambda)), \tag{4.1}$$

$$Q^0_\lambda(K^b) = \mathrm{Id}\,. \tag{4.2}$$

We construct K^b as an asymptotic sum of

$$K_j \in I^{-\frac{2n-1}{4} + j + i\alpha, -\frac{1}{4} - i\alpha}_{sc}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2}).$$

We would like

$$(\Delta_X - \lambda^2 + V)K_0 \in I^{-\frac{2n-1}{4} + i\alpha + 2, \frac{3}{4} - i\alpha}_{sc}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2})\ , \tag{4.3}$$

$$\sigma_0(Q^0_\lambda(K_0)) = \sigma_0(\mathrm{Id}); \tag{4.4}$$

then $(\Delta_X - \lambda^2 + V)K_0$ is automatically in

$$I_{sc}^{-\frac{2n-1}{4}+i\alpha+2,\frac{7}{4}-i\alpha}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2})$$

by Lemma 4.2. We also want

$$(\Delta_X - \lambda^2 + V)\left(\sum_{l=0}^{j-1} K_l\right) \in I_{sc}^{-\frac{2n-1}{4}+i\alpha+j+2,\frac{3}{4}-i\alpha}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2})$$

and this, of course, implies that it will also be an element of

$$I_{sc}^{-\frac{2n-1}{4}+i\alpha+j+2,\frac{7}{4}-i\alpha}(X \times \partial X, \tilde{G}(\lambda); {}^{sc}\Omega^{1/2}).$$

Now near $G \cap G^\sharp(\lambda)$, where G is smooth, we can as in [5] give an explicit construction, and it is then necessary only to have that the principal symbol of $Q_\lambda^0(K_0)$ is equal to 1 to ensure that $Q_\lambda^0(K^b) = \mathrm{Id}$. We look for K_j of the form

$$x^{j+i\alpha}e^{i\lambda\phi(y,y')/x}a_j(x,y,y',\lambda)v, a_j \in C^\infty(X \times \partial X)$$

with v a fixed scattering half-density and ϕ the cosine of the Riemannian distance from y to y. Let $a_j^{'}$ be the restriction of a_j to $x = 0$. Taking geodesic normal coordinates, y, about each y' the transport equations for $a_j^{'}$ is of the form

$$(y \cdot \partial_y + j)a_j^{'} + b_j a_j^{'} = c_j \in C^\infty(X \times \partial X)$$

near $y = 0$, where c_0 is identically zero and b_j vanishes quadratically at $y = 0$. Note that the extra power $x^{i\alpha}$ cancels with the extra term coming from the lead term of the potential. Thus, as in [5], the terms K_j exist sc-microlocally close to $G^\sharp(\lambda)$.

We now need to continue each K_j up to $G^\sharp(-\lambda)$; we do so by solving transport equations for the principal symbols and iteratively solving away the error.

Now the principal symbol of K_0, $\sigma_m(K_0)$ is of the form

$$b\frac{|ds|^{\frac{1}{2}}|dy|^{\frac{1}{2}}|d\hat{\mu}|^{\frac{1}{2}}}{(\sin s)^{\frac{1}{2}}}|dx|^{m-\frac{2n-1}{4}}.$$

On the lifted geodesic $\beta(s)$ the subprincipal term $c(\beta(s)) = 2\lambda\sin(s)d(\beta(s))$ for some smooth d. From Proposition 3.2, the transport equation for b is

$$\frac{2\lambda}{i}\left(\sin(s)\frac{d}{ds} + \left(\frac{1-n}{2} + i\alpha\right)\cos(s) + i\sin(s)d(\beta(s))\right)b + Ab = 0$$

with the final term being the contribution coming from the long range nature of the potential. Writing $\tilde{b} = (e^{i\int d(\beta(s'))ds'}\sin(s))^{\frac{1-n}{2}+i\alpha}b$, we thus have

$$\frac{d\tilde{b}}{ds} + i\frac{\alpha}{\sin(s)}\tilde{b} = 0$$

(using the fact that $\alpha = V/(2\lambda)$). Introducing an integrating factor, this becomes

$$\frac{d}{ds}\left[e^{i\alpha\int\frac{1}{\sin(s')}ds'}\tilde{b}\right] = 0.$$

This means that

$$\begin{aligned} b = &\ C\sin(s)^{\frac{n-1}{2}-i\alpha}e^{-i\alpha\int\frac{1}{\sin(s')}ds'}e^{-i\int d(\beta(s'))ds'} \\ = &\ C\sin(s)^{\frac{n-1}{2}-i\alpha}\tan(s/2)^{-i\alpha}e^{-i\int d(\beta(s'))ds'}. \end{aligned}$$

As $s\to\pi-$, that is, near $G^\sharp(\lambda)$, this has a singularity of the form $(\pi-s)^{\frac{n-1}{2}}$, and as $s\to 0+$, this has the form $s^{\frac{n-1}{2}-2i\alpha}$.

As the order on $G^\sharp(\lambda)$ is $p_1=-\frac{1}{4}+i\alpha$ and on $G^\sharp(-\lambda)$ is $p_2=-\frac{1}{4}-i\alpha$, the orders p_1-m and p_2-m are equal to the orders of singularities, and thus the solution of the transport equation is a legitimate symbol and we can construct K_0.

Now by Lemma 4.2,

$$(\Delta+V-\lambda^2)(K_0)\in I^{m+2,p_1+2,p_2+2}(X\times\partial X,\tilde{G};{}^{sc}\Omega^{1/2}),$$

and we look for

$$K_1\in I^{m+1,p_1,p_2}(X\times\partial X,\tilde{G};{}^{sc}\Omega^{1/2}),$$

such that

$$(\Delta+V-\lambda^2)(K_0+K_1)\in I^{m+3,p_1+1,p_2+1}(X\times\partial X,\tilde{G};{}^{sc}\Omega^{1/2}).$$

Now by Lemma 4.2, we have $I^{m+3,p_1+2,p_2+2}(X\times\partial X,\tilde{G};{}^{sc}\Omega^{1/2})$. Letting the principal symbol of K_1 be $b_1|dx|^{m+1-\frac{2n-1}{4}}$ times the trivializing density above, we obtain a transport equation; arguing as above, it becomes

$$\begin{aligned} &\sin(s)^{\frac{n-1}{2}+i\alpha}\tan(s/2)^{-i\alpha}e^{-i\int d(\beta(s'))ds'} \\ &\qquad \frac{d}{ds}e^{i\int d(\beta(s'))ds'}\left(\sin(s)^{\frac{1-n}{2}+1+i\alpha}\tan(s/2)^{i\alpha}b_1\right) \\ &\qquad\qquad = g(s)e^{-i\int d(\beta(s'))ds'}\sin(s)^{\frac{n-1}{2}-i\alpha}\tan(s/2)^{-i\alpha} \end{aligned}$$

with $g(s)$ a smooth function on $[0,\pi]$ (and depending smoothly on the suppressed parameters). Canceling, we obtain that

$$\frac{d}{ds}\left(e^{i\int d(\beta(s')ds'}\sin(s)^{\frac{1-n}{2}+1+i\alpha}\tan(s/2)^{i\alpha}b_1\right)=g(s),$$

which has a solution in the appropriate symbol class. The same argument, after appropriately shifting indices, constructs all the terms K_j.

Asymptotically summing, we obtain K^b such that

$$(\Delta+V-\lambda^2)K^b\in I^{7/4+i\alpha}(G^\sharp(\lambda))+I^{7/4-i\alpha}(G^\sharp(-\lambda)).$$

These errors can now be removed by an iterative construction of their Taylor series (cf. Lemma 16 of [5]), and we obtain K as desired. □

The remainder of the proof of Theorem 1.1 is identical to that of the main theorem in [5] and we therefore omit the details.

5. Recovering asymptotics of potentials. In this section, we prove Theorem 1.2 and recall from [3] the deduction of Theorem 1.3. Note that the lead term of the Coulomb-like potential is automatically determined by the order of the elliptic Fourier integral operator $S(\lambda)$. Thus we need consider only potentials which have the same lead term and thus which have Poisson operators with the same principal symbol.

Let V_1, V_2 be two Coulomb-like potentials with lead term A such that

$$V_1 - V_2 = x^k W + O(x^{k+1})$$

with $W \in C^\infty(\partial X)$. Letting P_1, P_2 be the associated Poisson operators at energy λ, we then have that

$$(\Delta - \lambda^2)P_j = -V_j P_j$$

and thus that

$$(\Delta - \lambda^2)(P_1 - P_2) = -V_1 P_1 + V_2 P_2 = -V_1(P_1 - P_2) + (V_2 - V_1)P_2.$$

Hence

$$(\Delta + V_1 - \lambda^2)(P_1 - P_2) = (V_2 - V_1)P_2.$$

We also have that

$$Q^0_{-\lambda}(P_1 - P_2) = \mathrm{Id} - \mathrm{Id} = 0.$$

Thus as in [3], we conclude that $P_1 - P_2$ is of order $-\frac{2n-1}{4} + i\alpha - k$ on G as the transport equations for the principal symbol at each level will be homogeneous with zero initial data and thus have zero solution. At the kth level, the equation becomes inhomogeneous, and we obtain the following transport equation, using Proposition 4.1 and Proposition 3.2:

$$\frac{2\lambda}{i}\left(\sin(s)\frac{d}{ds} + \left(\frac{1-n}{2} + k + i\alpha\right)\cos(s) + i\sin(s)d(\beta(s))\right)b + Ab$$
$$= CW(\gamma(s))e^{-i\int d(\beta(s'))ds'}\sin(s)^{\frac{n-1}{2}-i\alpha}\tan(s/2)^{-i\alpha},$$

where $b|dx|^{-\frac{2n-1}{2}+i\alpha+k}\frac{|ds|^{\frac{1}{2}}|dy|^{\frac{1}{2}}|d\hat{\mu}|^{\frac{1}{2}}}{(\sin s)^{\frac{1}{2}}}$ is the principal symbol of $P_1 - P_2$. Writing $\tilde{b} = e^{i\int d(\beta(s'))ds'}(\sin(s))^{\frac{1-n}{2}+k+i\alpha}b$, this becomes

$$\frac{d\tilde{b}}{ds} = \frac{i}{2\lambda}(\sin(s))^{k-1}W(\gamma(s)),$$

where $\gamma(s)$ is the geodesic in ∂X. Thus as in [3], on approach to $G^\sharp(-\lambda)$ the symbol will be $(\sin(s))^{\frac{n-1}{2}-k-i\alpha}\tan(\frac{s}{2})^{-i\alpha}e^{-i\int d(\beta(s'))ds'}$ times the weighted integral

$$\int_0^\pi (\sin(s))^{k-1}W(\gamma(s))ds$$

and thus from Proposition 3.3, the principal symbol of

$$S_1 - S_2 = Q^0_\lambda(P_1 - P_2)$$

will be a fixed elliptic factor times $\int_0^\pi(\sin(s))^{k-1}W(\gamma(s))ds$; the result follows. Note that the geodesic γ required to compute $\sigma_{1-k-2i\alpha}(S_1 - S_2)$ at a point (x,ξ) is just the geodesic, which when lifted, ends at (x,ξ).

Acknowledgments. I would like to thank Antonio Sa Barreto and Maciej Zworski for helpful conversations. This paper was prepared while participating in the microlocal methods programme at the Fields Institute in the fall of 1997, and I would like to thank the Fields Insititute and the progamme organizers for their hospitality.

REFERENCES

[1] V. Enss and R. Weder, *The geometrical approach to multidimensional inverse scattering,* J. Math. Phys., 36 (1995), pp. 3902–3921.
[2] H. Isozaki and H. Kitada, *Scattering matrices for two body Schrödinger operators,* Sci. Papers College Arts Sci. Univ. Tokyo, 35 (1986), pp. 81–107.
[3] M.S. Joshi and A. Sa Barreto, *Recovering asymptotics of short range potentials*, Comm. Math. Phys., 193 (1998), pp. 197–208.
[4] R.B. Melrose, *Spectral and scattering theory for the Laplacian on asymptotically Euclidean spaces*, in Spectral and Scattering Theory, M. Ikawa, ed., Marcel Dekker, New York, 1994.
[5] R.B. Melrose and M. Zworski, *Scattering metrics and geodesic flow at infinity*, Invent. Math., 124 (1996), pp. 389–436.
[6] R. Newton, *Construction of potentials from the phase shifts at fixed energy*, J. Math. Phys., 3 (1962), pp. 75–82.
[7] T. Regge, *Introduction to complex orbital moments*, Nuovo Cimento, 14 (1959), pp. 951–976.
[8] P. Sabatier, *Asymptotic properties of the potentials in the inverse scattering problem at fixed energy*, J. Math. Phys., 7 (1966), pp. 1515–1531.
[9] A. Vasy, *Geometric scattering theory for long range potentials and metrics*, Internat. Math. Res. Notices, (1998), pp. 285–315.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 527–536

ON THE ATTAINABLE EIGENVALUES OF THE LAPLACE OPERATOR*

DORIN BUCUR†, GIUSEPPE BUTTAZZO‡, AND ISABEL FIGUEIREDO§

Abstract. We consider the subset E of $\mathbb{R}^2$ of all points whose first and second components, respectively, coincide with the first and second eigenvalues of the Laplace operator $-\Delta$ with zero boundary conditions on domains of $\mathbb{R}^N$ with prescribed measure. We show that the set E is closed in $\mathbb{R}^2$.

Key words. shape optimization, eigenvalues, Dirichlet–Laplace operator

AMS subject classifications. 35P99, 49J20, 49Q10

PII. S0036141097329135

1. Introduction.

Let $B \subseteq \mathbb{R}^N$ be a fixed ball and let $c > 0$ be a positive number. We denote by

$$\mathcal{A}_c(B) = \big\{ A \subseteq B : \quad A \quad \text{quasi open, } m(A) \le c \big\}$$

the family of all quasi-open subsets of B having Lebesgue measure less than or equal to c, and by $s : \mathcal{A}_c(B) \to \mathbb{R}^2$ the "spot" function defined by

$$s(A) = (\lambda_1(A), \lambda_2(A)),$$

where $\lambda_1(A), \lambda_2(A)$ are the first two eigenvalues (counted with their multiplicities) of the Laplace operator $-\Delta$ on the Sobolev space $H_0^1(A)$. Since we will always work with Sobolev functions it is convenient to consider domains which are quasi-open (that is, subsets of $\mathbb{R}^N$ of the form $\{u > 0\}$ where $u \in H^1(\mathbb{R}^N)$) instead of usual domains which are open subsets of $\mathbb{R}^N$. This choice is also justified by the relaxation theory of Dirichlet problems. In fact, the family of quasi-open sets is the largest class of domains contained in the closure of the family of open sets with respect to the γ-topology, on which Sobolev spaces H_0^1 are well defined.

The purpose of this paper is to prove that the range of s is closed in $\mathbb{R}^2$ if, for a given c, the ball B is large enough. This will immediately imply the existence of a solution for problems of the form

$$\min \Big\{ \Phi(\lambda_1(A), \lambda_2(A)) : \quad A \in \mathcal{A}_c(B) \Big\}$$

for a large class of cost functions Φ.

*Received by the editors October 13, 1997; accepted for publication (in revised form) May 26, 1998; published electronically March 30, 1999.

http://www.siam.org/journals/sima/30-3/32913.html

†CNRS-Equipe de Mathématiques, Université de Franche-Comté, 16 route de Gray, 25030 Besançon, France (bucur@math.univ-fcomte.fr). This author was supported by the ESF/FBP project on Mathematical Treatement of Free Boundary Problems and by the University of Pisa.

‡Dipartimento di Matematica, Università di Pisa, Via Buonarroti 2, 56127 Pisa, Italy (buttazzo@dm.unipi.it). The work of this author is part of the EEC/HCM project "Phase Transition Problems and Singular Perturbations," contract CHRX-CT94-0608.

§Departamento de Matemática, Universidade de Coimbra, Apartado 3008, 3000 Coimbra, Portugal (isabelf@mat.uc.pt). The work of this author is part of the EEC/HCM project "Shells: Mathematical Modeling and Analysis, Scientific Computing," contract ERBCHRX-CT94-0536.

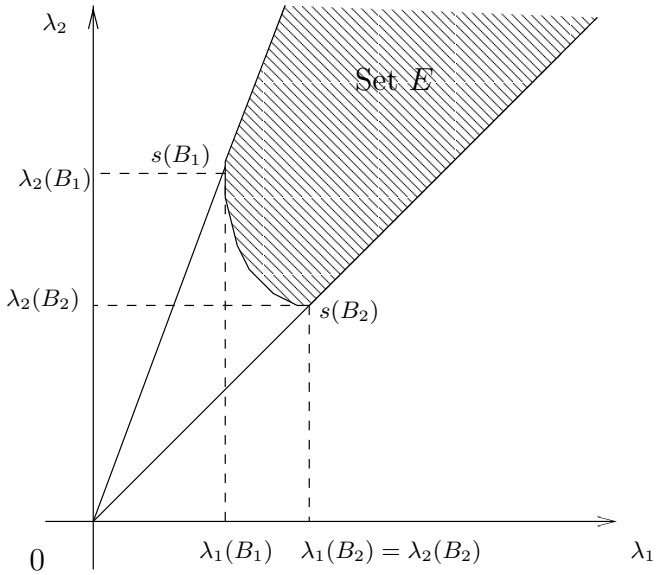


FIG. 1.

Let us denote by $E = s(\mathcal{A}_c(B))$ the image of s in $\mathbb{R}^2$. Some classical remarks already give an idea where the set E lies.

Denoting by B_1 the ball of mass c and by B_2 the union of two disjoint balls of mass $\frac{c}{2}$, classical arguments give that

- $\forall A \in \mathcal{A}_c(B), \quad \lambda_1(A) \geq \lambda_1(B_1)$ (proved by Faber [10] and Krahn [15], [16]);
- $\forall A \in \mathcal{A}_c(B), \quad \lambda_2(A) \geq \lambda_2(B_2) = \lambda_1(B_2)$ (we refer to Krahn [16]; see also [17] for a proof by P. Szegö);
- $\forall A \in \mathcal{A}_c(B), \quad \lambda_2(B_1)/\lambda_1(B_1) \geq \lambda_2(A)/\lambda_1(A) \geq 1$ (recently proved by Ashbaugh and Benguria (see [1]));
- the set E is conical with respect to the origin, that is $(tx, ty) \in E$ whenever $(x, y) \in E$ and $t \geq 1$ (by homothety of ratio $\frac{1}{\sqrt{t}}$).

For a numerical study of the set E in the case $N = 2$ we refer the interested reader to the paper by Wolf and Keller (see [18]), where the picture for E (see Figure 1) is obtained.

Unfortunately, we are not able to prove the convexity of the set E, which the picture above seems to show; this would imply the closure result quite straightforwardly. However, we can prove that E is convex horizontally and vertically, and this is enough to imply that it is closed (see Lemma 2.1).

In the paper we often use the definition of γ-convergence for sequences of domains and some of its properties; this definition, which we recall in section 2, was introduced by Dal Maso and Mosco (see [9]) and turned out to be a very powerful tool in several shape optimization problems (see Buttazzo and Dal Maso [7], [8]).

We conclude the paper with a section where some open problems are presented.

2. The main theorem and some preliminary results. Let $c > 0$ be given and let $B \subseteq \mathbb{R}^N$ be a ball containing two disjoint balls of mass $\frac{c}{2}$. We shall prove the following result.

THEOREM 2.1. *The set E is closed in $\mathbb{R}^2$.*

The proof of the theorem above is based on the following lemma.

LEMMA 2.1. *If the set E is convex on the vertical and horizontal directions, then E is closed in $\mathbb{R}^2$.*

Following this lemma it suffices to prove the convexity of E on vertical and horizontal directions. For this purpose, we shall split the proof in two steps:

Step 1. A is convex on horizontal lines; namely, if $A \in \mathcal{A}_c(B)$, then the segment joining $(\lambda_1(A), \lambda_2(A))$ to $(\lambda_2(A), \lambda_2(A))$ is contained in E.

Step 2. A is convex on vertical lines; namely, if $A \in \mathcal{A}_c(B)$, then the segment joining $(\lambda_1(A), \lambda_2(A))$ to $(\lambda_1(A), \frac{\lambda_2(B_1)}{\lambda_1(B_1)}\lambda_1(A))$ is contained in E.

The proofs of Lemma 2.1 and of Steps 1 and 2 will be given in section 3; we recall now classical notions and give some preliminary results.

The capacity of a set $E \subseteq B$ is defined by

$$cap(E) = \inf\left\{\int_B |\nabla u|^2, \quad u \in \mathcal{U}_E\right\},$$

where $\mathcal{U}_E$ is the class of all functions $u \in H_0^1(B)$ such that $u \geq 1$ almost everywhere (a.e.) in a neighborhood of E. We say that a property $p(x)$ holds quasi everywhere on E (q.e. on E) if the set of all points $x \in E$ for which $p(x)$ does not hold has capacity zero.

A subset $A \subseteq B$ is called quasi-open if for every $\epsilon > 0$ there exists an open subset G_ϵ of B such that $A \cup G_\epsilon$ is open and $cap(G_\epsilon) < \epsilon$. It easily can be seen that for any quasi-open set there exists a decreasing sequence $\{A_n\}_{n\in N}$ of open sets containing A such that $cap(A_n \setminus A) \to 0$. A function $f : B \mapsto \mathbb{R}$ is said to be quasi-continuous if for all $\epsilon > 0$ there exists an open set G_ϵ with $cap(G_\epsilon) < \epsilon$ such that f is continuous on $B \setminus G_\epsilon$ (see [13], [19]). For a quasi-open set A, the Sobolev space $H_0^1(A)$ is defined as

$$H_0^1(A) = \{u \in H_0^1(B) \ : \quad u = 0 \quad \text{q.e. on} \quad B \setminus A\}.$$

The fine topology on B is the coarsest topology making all superharmonic functions continuous. The relation between the quasi topology and the fine topology is studied in [12], [14].

For a quasi-open set $A \in \mathcal{A}_c(B)$, we denote by $\lambda_1(A)$, $\lambda_2(A)$ the first two eigenvalues (counted with their multiplicity) of the Laplace operator $-\Delta$ on $H_0^1(A)$. They are given by the classical formulae

$$\lambda_1(A) = \min_{\substack{\varphi\in H_0^1(A)\\ \varphi\neq 0}} \frac{\int_A |\nabla\varphi|^2 dx}{\int_A |\varphi|^2 dx},$$

$$\lambda_2(A) = \max_{\substack{\psi\in H_0^1(A)\\ \psi\neq 0}} \min_{\substack{\varphi\in H_0^1(A)\setminus\{0\}\\ \varphi\perp\psi}} \frac{\int_A |\nabla\varphi|^2 dx}{\int_A |\varphi|^2 dx}.$$

If we denote by φ_1, φ_2 the eigenfunctions corresponding to $\lambda_1(A)$ and $\lambda_2(A)$, then we have

$$\begin{aligned} &\varphi_1 \in H_0^1(A) \quad \text{and} \quad -\Delta\varphi_1 = \lambda_1(A)\varphi_1 \quad \text{in} \quad H_0^1(A),\\ &\varphi_2 \in H_0^1(A) \quad \text{and} \quad -\Delta\varphi_2 = \lambda_2(A)\varphi_2 \quad \text{in} \quad H_0^1(A). \end{aligned}$$

Generally, we shall use the notation $A_1 = \{\varphi_1 > 0\}$ and $A_2 = \{\varphi_2 \neq 0\}$, where we denote here by φ_1 and φ_2 the quasi-continuous representatives of the corresponding eigenfunctions. In order to establish the possible relations between A_1 and A_2 we shall give a lemma which is an extension of a classical result on open sets.

LEMMA 2.2. *If C_1, C_2 are two quasi-open sets with $cap(C_1 \cap C_2) = 0$ and $u \in H_0^1(C_1 \cup C_2)$, then $u_{|C_1} \in H_0^1(C_1)$ and $u_{|C_2} \in H_0^1(C_2)$.*

Proof. There exists a sequence of elements $u_n \in \mathcal{D}(C_1 \cup C_2)$ such that $u_n \to u$ strongly in $H_0^1(C_1 \cup C_2)$. (See [14] for the definition of $\mathcal{D}(C_1 \cup C_2)$.) Therefore it suffices to consider only elements of $\mathcal{D}(C_1 \cup C_2)$ since $u_{n|C_1} \to u_{|C_1}$ and $u_{n|C_2} \to u_{|C_2}$. Following Lemma 2.4 of [14], there exists a sequence of functions $g_k \to u$ strongly in $H_0^1(C_1 \cup C_2)$ and $g_k = g_k^1 + g_k^2$ where $g_k^1 \in H_0^1(C_1)$ and $g_k^2 \in H_0^1(C_2)$ since $\{C_1, C_2\}$ is a quasi covering of $C_1 \cup C_2$. Since $cap(C_1 \cap C_2) = 0$ the functions g_k^1 and g_k^2 are orthogonal in $L^2(B)$ and in $H_0^1(B)$; this implies that the sequences $\{g_k^1\}_k, \{g_k^2\}_k$ are bounded in both spaces. For subsequences still denoted with the same index we get $g_k^1 \rightharpoonup u^1$ weakly in $H_0^1(C_1)$ and $g_k^2 \rightharpoonup u^2$ weakly in $H_0^1(C_2)$. From the strong L^2 convergence we have $u^1 = u_{|C_1}$ and $u^2 = u_{|C_2}$, which concludes the proof. □

We can formulate now the following lemma.

LEMMA 2.3. *Let A be a quasi-open set such that $\lambda_1(A) < \lambda_2(A)$. Then the fine interior of A_1 is finely connected and there are two possibilities: either $A_2 \subseteq A_1$ or $cap(A_1 \cap A_2) = 0$ for a convenient second eigenfunction φ_2.*

Proof. If A is open the result is immediate. If A is quasi open, the proof is similar and based on the previous lemma and the following assertion (see [12]): any positive superharmonic function on a finely open and connected set is either strictly positive or equal to zero. Particularly, this will be the case of the first eigenfunction.

Indeed, if A_1 is not finely connected (we denoted here by A_1 its fine interior), then it can be decomposed in a union of disjoint finely connected components $\{C_i\}_{i\in I}$ and, since $\varphi_{1|C_i} \in H_0^1(C_i) \subseteq H_0^1(A)$, we have that

$$\forall i \in I \qquad \frac{\int_{C_i} |\nabla \varphi_{1|C_i}|^2 dx}{\int_{C_i} |\varphi_{1|C_i}|^2 dx} = \lambda_1(A).$$

Thus, if I contains at least two indices this would mean that $\lambda_1(A)$ is at least double since we have two independent eigenfunctions (defined by the restriction of φ_1 on each set). Therefore A_1 has only one finely connected component.

Suppose now that $cap(A_1 \cap A_2) \neq 0$. Decomposing $A_2 = \cup_{i\in I} C_i'$, C_i' being finely connected, then for any component for which $cap(A_1 \cap C_i') \neq 0$ we have $C_i' \subseteq A_1$; otherwise $C_i' \cup A_1$ would be finely connected and φ_1 could not vanish on $C_i' \setminus A_1$. Thus the finely connected components of A_2 are of two types, $C_i' \subseteq A_1$ and $cap(C_j' \cap A_1) = 0$. In this case we can see that $\varphi_{2|\cup C_i'}$ and $\varphi_{2|\cup C_j'}$ are both orthogonal to φ_1 and they are still second eigenfunctions. Then A_2 can be chosen as $\cup C_i'$ or $\cup C_j'$. □

In order to introduce the γ-convergence for a sequence $\{A_n\}_{n\in N}$ in $\mathcal{A}_c(B)$ we recall the Mosco conditions:

(M_1) $\forall \varphi \in H_0^1(A)$, $\exists \varphi_n \in H_0^1(A_n)$, such that $\varphi_n \to \varphi$ strongly in $H_0^1(B)$.

(M_2) $\forall \varphi_{n_k} \in H_0^1(A_{n_k})$, with $\varphi_{n_k} \rightharpoonup \varphi$ weakly in $H_0^1(B)$, we have $\varphi \in H_0^1(A)$.

It is said that A_n γ-converges to A if M_1 and M_2 hold simultaneously (see, for instance, [8]). It is said that A_n weakly γ-converges to A if $A = \{w > 0\}$ and $w_n \rightharpoonup w$ weakly in $H_0^1(B)$, where w_n are the solutions of

$$\begin{cases} -\Delta w_n = 1, \\ w_n \in H_0^1(A_n) \end{cases}$$

extended by zero on $B \setminus A_n$. It is known that the weak γ-convergence is sequentially compact on $\mathcal{A}_c(B)$ (see [5]).

An important tool that we shall use in the proof of Steps 1 and 2 is the continuous Steiner symmetrization (CSS) (see Brock [3], [4]). It is well known that the CSS of an open set keeps constant its measure and decreases the first Dirichlet eigenvalue of the Laplacian. Here, we are interested in the behavior of the second eigenvalue with respect to the CSS, which derives from the Mosco behavior of the associated Sobolev space (see [6]). Let's recall from [6] some useful results. Consider a measurable set A and a hyperplane $\mathcal{H} \subseteq \mathbb{R}^N$. For $t \in [0,1]$ denote by A^t the Steiner symmetrization of A at time t in the orthogonal direction to $\mathcal{H}$.

PROPOSITION 2.1. *If A is open, the mapping $t \mapsto A^t$ is γ-continuous from the left and M_1 continuous from the right.*

For a quasi-open set $A \in \mathcal{A}_c(B)$, the set A^t is defined in the following way: Consider a decreasing sequence of open sets $\{A_n\}_{n\in N}$ with $cap(A_n \setminus A) \to 0$ and $A \subseteq A_n \subseteq B$. For any $t \in [0,1]$ the set A_n^t is well defined and by monotonicity we define $A_n^t \supseteq A_{n+1}^t$. Then $\{A_n^t\}_{n\in N}$ is γ-convergent and

$$A^t = \gamma - \lim_{n\to\infty} A_n^t.$$

PROPOSITION 2.2. *If A is quasi-open, the mapping $t \mapsto A^t$ is M_2 continuous from the left and M_1 continuous from the right.*

In terms of eigenvalues, Proposition 2.2 gives the following corollary.

COROLLARY 2.1. *For every $A \in \mathcal{A}_c(B)$ and every positive integer i, the mappings $t \mapsto \lambda_i(A^t)$ are lower-semicontinuous on the left and upper-semicontinuous on the right.*

This result will permit us to prove that the second eigenvalue has a Darboux-like property, that is, if $t_1 < t_2$ and $\lambda_2(A^{t_1}) < \lambda_2(A^{t_2})$, then $\forall \lambda^* \in [\lambda_2(A^{t_1}), \lambda_2(A^{t_2})]$ there exists $t^* \in [t_1, t_2]$ such that $\lambda_2(A^{t*}) = \lambda^*$.

In the proof of Steps 1 and 2, the idea is to make a sequence of CSS to transform a given quasi-open set $A \in \mathcal{A}_c(B)$ into a ball. Here, one can see that the choice of B is important since if $A \in \mathcal{A}_c(B)$ for any hyperplane $\mathcal{H}$ we still have $A^t \in \mathcal{A}_c(B)$.

For a compact set $K \in \mathbb{R}^N$ it is known the existence of a sequence of hyperplanes $\{\mathcal{H}_n\}_{n\in N}$ such that, denoting $K_0 = K$ and K_n the symmetrization of K_{n-1} with respect to $\mathcal{H}_n$, we have $m(K_n \Delta K^{\#}) \to 0$ (generally by $C^{\#}$ we denote the closed ball of measure $m(C)$; see [2]). If the convergence in measure is replaced by the Hausdorff convergence, a similar type of result can be found in Federer (see [11]).

For quasi-open sets we can formulate the following.

PROPOSITION 2.3. *Let $A \in \mathcal{A}_c(B)$. There exists a sequence of Steiner symmetrizations of A, denoted $\{A_n\}_{n\in N}$, such that $m(A_n \setminus A^{\#}) \to 0$ for $n \to \infty$.*

Proof. This result appears to be weaker than the similar one for compact sets, but nevertheless it is still sufficient for the proof of Step 2.

Suppose first that A is open. Consider $K_1 \subset\subset A$ such that $m(A \setminus K_1) \leq \epsilon_1/2$. We make a finite number of Steiner symmetrizations given by the result of [2] for K_1 such that

$$m\Big((K_1)_{n_1} \Delta K_1^{\#}\Big) \leq \frac{\epsilon_1}{2}.$$

Then, by monotonicity,

$$m\Big(A_{n_1} \setminus A^{\#}\Big) \leq \frac{\epsilon_1}{2} + \frac{\epsilon_1}{2} = \epsilon_1.$$

Choosing now another set $K_2 \subset\subset A_{n_1}$ with $m(A_{n_1} \setminus K_2) \le \epsilon_2/2$ we continue the process and obtain

$$m\Big((K_2)_{n_2} \setminus K_2^{\#}\Big) \le \frac{\epsilon_2}{2},$$

and so on. Choosing a sequence $\epsilon_n \to 0$, we conclude the proof in the case of open sets.

If A is quasi-open, consider a sequence of open sets $\{C_r\}_{r\in N}$ such that

$$A \subseteq C_{r+1} \subseteq C_r \subseteq B$$

and $cap(C_r \setminus A) \to 0$. We apply the previous result for C_r and begin by making a finite number of symmetrizations to C_1 such that

$$m\Big((C_1)_{n_1} \setminus C_1^{\#}\Big) \le \epsilon_1.$$

Then $m(A_{n_1} \setminus C_1^{\#}) \le \epsilon_1$. Now making a finite number of symmetrizations for C_2 we get $m((C_2)_{n_2} \setminus C_2^{\#}) \le \epsilon_2$, and so on. Finally we get $m(A_n \setminus A^{\#}) \to 0$, since $m(C_n^{\#} \Delta A^{\#}) \to 0$. □

COROLLARY 2.2. *For every $A \in \mathcal{A}_c(B)$ there exists a sequence $\{A_n\}_{n\in N}$ of Steiner symmetrizations of A such that any weak γ-limit point of $\{A_n\}_{n\in N}$ is contained in $A^{\#}$.*

Proof. Indeed, from the previous proposition we have $m(A_n \setminus A^{\#}) \to 0$. If U is the weak γ-limit of $\{A_{n_k}\}$, then $w_{n_k} \rightharpoonup w$ weakly in $H_0^1(B)$ and $U = \{w > 0\}$. Since $m(A_n \setminus A^{\#}) \to 0$ and $w_{n_k} \to w$ in $L^2(B)$, we get $w = 0$ a.e. on $\mathbb{R}^N \setminus A^{\#}$, hence $w \in H_0^1(A^{\#})$, which means $U \subseteq A^{\#}$. □

COROLLARY 2.3. *For the sequence $\{A_n\}_{n\in N}$ given by Proposition 2.3 we have*

$$\lambda_2(A^{\#}) \le \liminf_{n\to\infty} \lambda_2(A_n).$$

3. Proof of the results. We proceed now with the proofs of Lemma 2.1 and of Steps 1 and 2. It is convenient to indicate by d_1 the half-line $\{ts(B_1) :\ t \ge 1\}$ and by d_2 the half-line $\{ts(B_2) :\ t \ge 1\} = \{(x,x) \in \mathbb{R}^2 :\ x \ge \lambda_1(B_2)\}$.

Proof of Lemma 2.1. Consider $(x,y) \in \bar{E}$. There exists a sequence of sets $\{A_n\}_{n\in N} \subseteq \mathcal{A}_c(B)$ such that $s(A_n) \to (x,y)$. From the weak γ-compactness of the set $\mathcal{A}_c(B)$ for a subsequence still denoted with the same indices, we can write $A_n \to A$ in the weak γ-sense. Then $A \in \mathcal{A}_c(B)$ and, since the eigenvalues of the Laplacian are weakly γ-lower-semicontinuous, we get $\lambda_1(A) \le \liminf_{n\to\infty} \lambda_1(A_n) = x$ and $\lambda_2(A) \le \liminf_{n\to\infty} \lambda_2(A_n) = y$. From the vertical convexity of E, the vertical segment joining $s(A)$ with the half-line d_1 is contained in E. If $y < \lambda_2(B_1)$, we can find the point $(\lambda_1(A), y)$ on this segment, and now using the horizontal convexity, the segment joining $(\lambda_1(A), y)$ to d_2 is in E. But this segment contains the point (x,y) since $\lambda_1(A) \le x$.

If $y \ge \lambda_2(B_1)$, then the horizontal convexity gives directly $(x,y) \in E$. □

We give now a general result which establishes the existence of a γ-continuous and decreasing homotopy between two quasi-open sets $A_1 \subseteq A_0$.

PROPOSITION 3.1. *Let $A_1 \subseteq A_0$ be two quasi-open sets. There exists a decreasing homotopy from A_0 to A_1 which is γ-continuous; namely, there exists a γ-continuous mapping $h : [0,1] \to \mathcal{A}(\mathbb{R}^N)$ such that for $t_1 < t_2$, $h(t_1) \supseteq h(t_2)$, and $h(0) = A_0$, $h(1) = A_1$.*

Proof. Denote by K a closed cube containing A_0. We shall divide the cube in 2^N equal closed cubes $K_0, \dots, K_{2^N-1}$; each cube K_i is analogously divided in 2^N closed cubes $K_{i0}, \dots, K_{i2^N-1}$, and so on. Then with each real number $t \in [0,1]$ written in the 2^N-basis by $0, \alpha_1, \alpha_2, \dots$ we associate the set

$$\Lambda_t = (A_0 \setminus F_t) \cup A_1,$$

where

$$F_t = \cup_{n=1}^{\infty} \cup_{i=0}^{\alpha_n - 1} K_{\alpha_1 \dots \alpha_{n-1} i}.$$

Remark first that Λ_t is quasi-open since F_t is quasi-closed. Indeed, let's denote by

$$F_{t,k} = \cup_{n=1}^{k} \cup_{i=0}^{\alpha_n - 1} K_{\alpha_1 \dots \alpha_{n-1} i}$$

the closed set consisting of the first k-blocks of F_t. Set also

$$\Lambda_{t,k} = (A_0 \setminus F_{t,k}) \cup A_1,$$

which is obviously quasi-open, and remark that

$$\cap_{k \geq 1} \Lambda_{t,k} = \Lambda_t.$$

Since $cap(\Lambda_{t,k} \setminus \Lambda_t) \to 0$ for $k \to \infty$ we get that Λ_t is quasi-open. Moreover, the mapping $t \to \Lambda_t$ is continuous in capacity. Indeed, fix $t \in [0,1]$ and consider $t_n \to t$. We have to distinguish two situations: either t has an infinite number of digits and is not finishing with $aa \dots aa \dots$, or t has a finite number of digits or finishes with $aa \dots aa \dots$ (a being the greatest digit in the basis 2^N, namely $a = 2^N - 1$). In the first case, if $t_n \to t$, then $\forall k \in \mathbb{N}\ \exists n_k \in \mathbb{N}$ such that $\forall n \geq n_k$, t_n, and t have the same first k digits. In this case

$$cap(A^t \Delta A^{t_n}) \leq cap(K_{\alpha_1 \dots \alpha_k}),$$

and we derive the continuity in capacity.

If t has a finite number of digits $t = 0.\alpha_1 \alpha_2 \dots \alpha_k$, then t written as

$$t = 0.\alpha_1 \alpha_2 \dots \alpha_k 0000 \dots$$

is identified with

$$t' = 0.\alpha_1 \alpha_2 \dots (\alpha_k - 1) aaaa \dots .$$

The difference between A^t and $A^{t'}$ is a point, hence of zero capacity. Consider $t_n \to t$. If $t_n \geq t$, the first k digits of t_n and t coincide for $n \geq n_k$. If $t_n < t$, then the first k digits of t_n and t' coincide for $n \geq n_k$ and the conclusion follows.

Since the mapping $t \mapsto \Lambda_t$ is obviously decreasing and γ-continuous, to achieve the proof it is enough to take

$$h(t) = \Lambda_t. \qquad \square$$

Proof of Step 1. Let $A \in \mathcal{A}_c(B)$. If there exists a subset A^* of A, such that $\lambda_1(A^*) = \lambda_2(A^*) = \lambda_2(A)$, then one can directly apply Proposition 3.1, and Step 1 is proved since there exists a decreasing and γ-continuous homotopy from A to A^*. Since

$\lambda_2(A) = \lambda_2(A^*)$, then by monotonicity $\lambda_2(\Lambda_t) = \lambda_2(A)$. Since the first eigenvalue is γ-continuous, for each $\alpha \in [\lambda_1(A), \lambda_2(A)]$ there exists some t_α such that $\lambda_1(\Lambda_{t_\alpha}) = \alpha$.

Let's prove now the existence of the set A^*. If $\lambda_1(A) = \lambda_2(A)$, there is nothing to prove. Hence we suppose $\lambda_1(A) < \lambda_2(A)$, and from what we have seen in the previous section we have two possibilities: either $A_2 = \{\varphi_2 \neq 0\} \subseteq \{\varphi_1 > 0\} = A_1$ (and in this case $A^* = A_2$) or $cap(A_1 \cap A_2) = 0$. Denoting by P_t the open half-space

$$P_t = \{(x_1, \dots, x_N) \in \mathbb{R}^N | t < x_1\},$$

there exists some $t_0 \in \mathbb{R}$ such that $\lambda_1(A_1 \cap P_{t_0}) = \lambda_2(A)$. Choosing

$$A^* = (A_1 \cap P_{t_0}) \cup A_2,$$

the conclusion follows. □

Proof of Step 2. Let's consider $A \in \mathcal{A}_c(B)$ and denote by $A^\#$ the closed ball of measure $m(A)$. The idea of proving the convexity in the vertical direction is to make a sequence of continuous Steiner symmetrizations transforming A such that $m(A_n \setminus A^\#) \to 0$ and to use the horizontal convexity. If $\lambda_2(A) \geq \lambda_2(A^\#)$, then the segment

$$\left\{ (\lambda_1(A), \gamma) \ : \quad \text{for} \quad \gamma \in \left[\lambda_2(A), \lambda_1(A) \frac{\lambda_2(B_1)}{\lambda_1(B_1)} \right] \right\}$$

is contained in E. This follows immediately from the convexity on the horizontal lines since all the half-line supported by d_1 and having B_1 as extreme point is in E.

So let's suppose $\lambda_2(A) < \lambda_2(A^\#)$ and choose $\alpha \in]\lambda_2(A), \lambda_2(A^\#)[$. We intend to prove that $(\lambda_1(A), \alpha) \in E$. We use Corollary 2.3 and we find a sequence of Steiner symmetrizations $\{A_n\}_{n \in N}$ such that $\liminf_{n \to \infty} \lambda_2(A_n) \geq \lambda_2(A^\#)$. In order to underline the evolution of the set A "toward" the ball, we say that the CSS from A_n to A_{n+1} is parametrized by $t \in [n, n+1]$ by simple translation of the interval $[0, 1]$. In this way we can define the set A^t for every $t \geq 0$, and the set A_n can also be written as A^n.

On the other hand, $\lambda_1(A_n) \leq \lambda_1(A)$. There exists some $n_0 \in N$ such that $\lambda_2(A_{n_0}) \geq \alpha$ and denote

$$t^* = \sup \Big\{ t \in [0, n_0] \ : \quad \lambda_2(A^t) \leq \alpha \Big\}.$$

From the upper-semicontinuity on the right we have $\lambda_2(A^{t^*}) \geq \alpha$ and from the lower-semicontinuity on the left we get $\lambda_2(A^{t^*}) \leq \alpha$ which give $\lambda_2(A^{t^*}) = \alpha$.

Using now the convexity on the horizontal lines, the segment joining $(\lambda_1(A^{t^*}), \alpha)$ with $(\alpha, \alpha) \in d_2$ is contained in E. However, since $\lambda_1(A^{t^*}) \leq \lambda_1(A)$, the point $(\lambda_1(A), \alpha)$ belongs to E. □

4. Further remarks. There are many other questions which can be raised. Is the set E convex? Is E still closed if the pair (λ_1, λ_2) is replaced by (λ_i, λ_j) or, more generally, if we consider the set

$$E_K = \Big\{ (\lambda_i(A))_{i \in K} : \quad A \in \mathcal{A}_c(B) \Big\},$$

where K is a given subset of positive integers? Are the sets A on the boundary of E smooth? If the ball B is replaced by an open set Ω, is the set $s(\mathcal{A}_c(\Omega))$ still closed? Or if the Laplace operator is replaced by

$$L = -\partial_i(a_{ij}\partial_j) + b_i\partial_i + c \ ?$$

Let's give a proposition which yields some information on the boundary of the set E. For a set $A \in \mathcal{A}_c(B)$ we shall denote

$$\mathcal{R}_{\inf}(A) = \{(x,y) \in \mathbb{R}^2 : x \le \lambda_1(A), y \le \lambda_2(A)\}.$$

PROPOSITION 4.1. *For every $A \in \mathcal{A}_c(B)$, there exists a set $\tilde{A} \in \mathcal{A}_c(B)$ which is either finely connected or two balls, such that $s(\tilde{A}) \in \mathcal{R}_{\inf}(A)$.*

Proof. Fix $A \in \mathcal{A}_c(B)$ and set $A_1 = \{\varphi_1 > 0\}$ and $A_2 = \{\varphi_2 \neq 0\}$. If $\lambda_1(A) = \lambda_2(A)$ the assertion is obvious since $s(A) \in d_2$. If $\lambda_1(A) < \lambda_2(A)$ there are two possibilities: if $A_2 \subseteq A_1$, then A_1 is finely connected and $s(A_1) \in \mathcal{R}_{\inf}(A)$; if $cap(A_2 \cap A_1) = 0$, we make the Schwarz rearrangements of A_1 and A_2 into the disjoint balls C_1 and C_2 and we get $s(C_1 \cup C_2) \in \mathcal{R}_{\inf}(A)$ and $C_1 \cup C_2 \in \mathcal{A}_c(B)$. □

This proposition means that any A whose $s(A)$ is on $\partial E \setminus (d_1 \cup d_2)$ is either finely connected or two balls. An open question is to study if these sets are simply connected.

To conclude, we remark that the result of this paper can be applied to prove the existence of solutions for some classes of shape optimization problems for which the shape functional is not monotone with respect with the set inclusion (see [8]). We can consider problems of the form

$$\min\Big\{\Phi(\lambda_1(A), \lambda_2(A)) : \quad A \in \mathcal{A}_c(B)\Big\}, \tag{4.1}$$

where $\Phi : E \to \overline{\mathbb{R}}$ is lower semicontinuous and goes to $+\infty$ at infinity. This is the case, for instance, of

$$\Phi(x,y) = (x-\alpha)^2 + (y-\beta)^2,$$

where (α, β) is any element in $\mathbb{R}^2$. Therefore by Theorem 2.1 the minimization problem (4.1) admits at least a solution.

Acknowledgments. The authors want to thank the Department of Mathematics of the University of Pisa, where this work was initiated, for the warm hospitality.

REFERENCES

[1] M. S. ASHBAUGH AND R. D. BENGURIA, *Proof of the Payne-Pólya-Weinberger conjecture*, Bull. Amer. Math. Soc., 25 (1991), pp. 19–29.
[2] H. J. BRASCAMP, E. LIEB, AND J. M. LUTTINGER, *A general rearrangement inequality for mulptiple integrals*, J. Funct. Anal., 17 (1974), pp. 227–237.
[3] F. BROCK *Continuous symmetrization and symmetry of solutions of elliptic problems*, Habilitation thesis, Leipzig, 1998.
[4] F. BROCK, *Continuous Steiner-symmetrization*, Math. Nachr., 172 (1995), pp. 25–48.
[5] D. BUCUR, G. BUTTAZZO, AND A. HENROT, *An existence result for some optimal partition problems*, Adv. Math. Sci. Appl., 8 (1998), pp. 571–579.
[6] D. BUCUR AND A. HENROT, *Stability for the Dirichlet Problem under Continuous Steiner Symmetrization*, Potential Anal., to appear.
[7] G. BUTTAZZO AND G. DAL MASO, *Shape optimization for Dirichlet problems: Relaxed formulation and optimality conditions*, Appl. Math. Optim., 23 (1991), pp. 17–49.
[8] G. BUTTAZZO AND G. DAL MASO, *An existence result for a class of shape optimization problems*, Arch. Rational Mech. Anal., 122 (1993), pp. 183–195.
[9] G. DAL MASO AND U. MOSCO, *Wiener's criterion and Γ-convergence*, Appl. Math. Optim., 15 (1987), pp. 15–63.
[10] G. FABER, *Beweis, dass unter allen homogenen Membranen von gleicher Fläche und gleicher Spannung die kreisförmige den tiefsten Grundton gibt*, Sitzungsber. Bayer. Akad. Wiss., (1923), pp. 169–172.

[11] H. Federer, *Geometric Measure Theory*, Springer-Verlag, Berlin, 1969.
[12] B. Fuglede, *Finely Harmonic Functions*, Lecture Notes in Math. 289, Springer-Verlag, Berlin, Heidelberg, New York, 1972.
[13] J. Heinonen, T. Kilpelainen, and O. Martio, *Nonlinear Potential Theory of Degenerate Elliptic Equations*, Clarendon Press, Oxford, New York, Tokyo, 1993.
[14] T. Kilpelainen and J. Maly, *Supersolutions to degenerate elliptic equations on quasi open sets*, Comm. Partial Differential Equations, 17 (1983), pp. 371–405.
[15] E. Krahn, *Über eine von Rayleigh formulierte Minimaleigenschaft des Kreises*, Math. Ann., 94 (1924), pp. 97–100.
[16] E. Krahn, *Über Minimaleigenschaften der Kugel in drei und mehr Dimensionen*, Acta Comm. Univ. Dorpat., A9 (1926), pp. 1–44.
[17] G. Polya, *On the characteristic frequencies of a symmetric membrane*, Math. Z., 63 (1955), pp. 331–337.
[18] S. A. Wolf and J. B. Keller, *Range of the first two eigenvalues of the Laplacian*, Proc. Roy. Soc. London Ser. A, 447 (1994), pp. 397–412.
[19] W. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 537–558

EXACT SOLUTIONS TO DEGENERATE CONSERVATION LAWS*

ROBIN YOUNG†

Abstract. We consider large variation solutions to systems of conservation laws, for which the Glimm–Lax theory of decay breaks down. We identify and isolate geometric nonlinearities which are distinct from the usual genuine nonlinearity of each wave field by describing some degenerate systems in which all nonlinearity is geometric and is manifested in the coupling of the different wave families. We then construct exact explicit solutions to these equations and examine properties of these solutions. We find a wide variety of phenomena, depending on the form of the nonlinearity. The most striking of these include strong nonlinear instability of solutions and nontrivial time-periodic solutions. We also find solutions which grow or decay exponentially and oscillating solutions which correspond to rotations by an irrational angle. These oscillating, periodic, and exponential solutions can all appear in a single system with small initial data, demonstrating sensitive dependence on initial conditions.

Key words. conservation laws, linear degeneracy, periodic solutions

AMS subject classification. 35

PII. S0036141097327239

1. Introduction.

We are interested in periodic solutions to systems of nonlinear conservation laws in one space dimension,

$$u_t + f(u)_x = 0,$$

where $u \in \mathbf{R}^N$ for $N \geq 3$. For these systems, the Glimm–Lax theory of decay breaks down due to geometric nonlinearities which are not present for systems of two equations.

The nonlinearity of the flux f is usually manifested as genuine nonlinearity: nonlinear wave-speeds lead to expansion of rarefactions and eventual shock formation. The appearance of shocks means that we must consider weak solutions, which are solutions of the integral conservation law

$$\int_0^\infty \int_{-\infty}^\infty (\psi_t u + \psi_x f(u))\, dx\, dt + \int_{-\infty}^\infty \psi(x,0) u_0(x)\, dx = 0,$$

where ψ is a smooth test function. After a shock forms, it interacts with rarefactions, leading to decay and loss of information. This is a stabilizing and time-irreversible effect. If the system consists of two equations, or if the total variation of the data is small, then this is the dominant phenomenon. Here the Glimm–Lax decay theory holds, and solutions decay like $1/t$ [2].

If the system consists of three or more equations and the total variation of the data is large, then a new nonlinear phenomenon emerges. This is a geometric nonlinear effect due to the coupling of the different wave families. This coupling is a

*Received by the editors September 12, 1997; accepted for publication (in revised form) July 23, 1998; published electronically March 30, 1999. This research was partially supported by DOE grant number DE-FG02-88ER25053 at the Courant Institute, and by NSF grant DMS-9201581 and DOE grant DE-FG02-90ER25084 at Stony Brook.

http://www.siam.org/journals/sima/30-3/32723.html

†Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA 01003 (young@math.umass.edu).

manifestation of the failure of the wave curves to form a coordinate system. The wave curves are composite rarefaction and shock curves and are approximately the integral curves of the eigenvector fields. The failure of these to commute is quantified by the Lie brackets

$$[r^j, r^k] = r^j \cdot \nabla r^k - r^k \cdot \nabla r^j = \sum \Lambda_i^{jk} r^i$$

of the eigenvectors. The *interaction coefficient* Λ_i^{jk} defined here represents the i-wave generated by the interaction of a j- and a k-wave, to leading order. These effects are quadratic and indeed yield all quadratic effects of interactions: when corrections for these terms are included, all errors become cubic [12].

These quadratic effects accumulate and become dominant when the total variation is large. Heuristically, this can be seen by counting waves and interactions: if all waves have strength $O(\epsilon)$ and there are approximately $O(V/\epsilon)$ waves, then there are $O(V^2/\epsilon^2)$ interactions. Thus the cumulative effect due to quadratic terms in interactions is $O(V^2)$, which does not vanish as the amplitude $\epsilon \to 0$. On the other hand, if these quadratic terms are not present, as is the case for two equations, the cumulative error becomes $O(V^2\epsilon)$ and can be controlled.

In this paper, we wish to investigate the consequences of these quadratic effects for solutions of large total variation, and in particular for data which is periodic. We study strictly hyperbolic systems in which there are no genuinely nonlinear fields, so that all the nonlinearity comes from the coupling of eigenvector fields. Specifically, we will assume that all wave-speeds, which are eigenvalues of the flux matrix $Df(u)$, are constant in a neighborhood. This assumption leads to simplifications which allow us to construct a class of exact weak solutions to the conservation law and study properties of these solutions. Since all fields are linearly degenerate, there are no shocks or rarefactions present, and all waves in our solutions are contact discontinuities. Moreover, the wave curves are then exactly the integrals of the eigenvector fields. We remark that in systems of two equations, the assumption of constant wave-speeds necessarily implies that the system be linear, as can be seen by rewriting the system in Riemann invariant coordinates.

The method of weakly nonlinear geometric optics has been used to predict a variety of new phenomena [9, 5, 6]. These include the strong nonlinear instability of solutions as well as the indefinite delay of shock formation. It is apparent from these simplified equations that the values of the interaction coefficients considered together are crucial to knowledge of the behavior of solutions. Thus we expect different behaviors depending on how many of these coefficients vanish and the positivity or negativity of others.

Not all of the interaction coefficients play an important role: with a careful normalization of the eigenvectors, all coefficients Λ_i^{ik} can be made to vanish at the origin [11]. Since the Lie bracket is antisymmetric, we thus need know only the values Λ_i^{jk}, where $j > k$ and both are distinct from i. Thus, for three equations we need only consider the triple $\{\Lambda_1, \Lambda_2, \Lambda_3\}$, where $\Lambda_1 = \Lambda_1^{32}$, etc. We shall see that different values for these triples lead to quite different qualitative behavior of solutions. This indicates that in order to make statements about the pointwise behavior of solutions, it will be necessary to make some assumptions about the systems and about the interaction coefficients in particular. Clearly, these assumptions should be based on physical principles.

The class of weak solutions which we study can be easily constructed due to the simplicity of the systems. We shall work with a 3×3 system for which the (constant)

wave-speeds are -1, 0, and 1. Our initial data is a piecewise constant with jumps at regular intervals. All waves are then contact discontinuities which propagate with speeds 0 or ± 1, and a repeating resonant interaction pattern develops. All interactions in this wave pattern are triple interactions and occur at the points of a regular lattice. In order to completely describe the solution, we need only keep track of the wave strengths. The problem thus reduces to a discrete problem, namely that of finding the wave strengths, which change only at points of interaction on the lattice.

We now describe some of the effects that appear for these systems depending on the choice of triples of interaction coefficients. The worst case of instability occurs when all three coefficients have the same sign: in this case a triple interaction may cause growth in each wave entering the interaction. Indeed, we shall set up a pattern in which each interaction has this magnification property, which has the effect of causing unbounded growth in amplitude of the solution in finite time. In this case, the wave strengths satisfy a Riccati-type difference equation,

$$\epsilon_{k+1} \approx \epsilon_k + \epsilon_k^2,$$

solutions of which explode in finite time. In fact, since our conservation law is completely linearly degenerate, the solutions can be scaled in such a way that the growth rate can be accurately controlled. We remark that the amplitudes do not actually become infinite, before which our methods break down, but become large relative to the initial amplitude. As a corollary, we find solutions which decay by a similar mechanism. We emphasize that this decay is not due to genuine nonlinearity and is the result of a different phenomenon. Indeed, all the solutions we consider are time-reversible, as may be expected from the lack of genuine nonlinearity.

This instability of solutions is the same phenomenon as that predicted by weakly nonlinear geometric optics in [5, 6]. However, here we are able to give an explicit description of the growth mechanism, as well as give accurate estimates of the rate of growth and the time of existence of solutions with sharp bounds on the total variation.

When considering realistic systems, physical assumptions lead to extra conditions on the interaction coefficients; the effects of various assumptions are described in [11]. The existence of a convex entropy function for which an additional conservation law can be derived is one such assumption and is equivalent to the assumption that we have a symmetric hyperbolic system. This symmetry leads to algebraic constraints on the interaction coefficients; for three equations, this is the relation

$$\frac{\Lambda_1}{\lambda_3 - \lambda_2} + \frac{\Lambda_2}{\lambda_3 - \lambda_1} + \frac{\Lambda_3}{\lambda_2 - \lambda_1} = 0.$$

In particular, this relation precludes the strong instability mentioned above, which requires that the Λ_i's have the same sign.

Another common constraint is the existence of a Riemann coordinate. This is a function whose gradient is a left eigenvector of the system, and implies that the corresponding interaction coefficient vanishes. This family can be thought of as being weakly decoupled from the system, in that it is not affected (to quadratic order) by the presence of waves of other families. We construct a class of degenerate systems of conservation laws with a 2-Riemann coordinateand study the variety of behaviors that can occur for these systems. In this class of examples, we are able to integrate the eigenvectors exactly and thus obtain exact formulæ for Riemann solutions and wave interactions. This in turn leads to a detailed description of the solutions constructed earlier. We remark that these solutions need not be small, and indeed the integrals of eigenvectors are globally defined.

Depending on the fluxes in these systems and on the initial data, our solutions exhibit qualitative differences. For a system with coefficients $\{1, 0, 1\}$, we find that we again obtain growing or decaying solutions, although in this case the rate of growth or decay is exponential. In particular, there is no finite time blowup in amplitude. Indeed, together with Blake Temple, the author has recently shown that solutions to systems of three conservation laws with one Riemann coordinate grow at most exponentially, so that this example shows that those results are sharp [10].

In [10], a new length scale was identified for solutions with large variation, which also determines the rate of growth of solutions. More precisely, a norm $\|u\|'_d$ is introduced which measures the variation of the solution over intervals of fixed size d, and the ratio $\rho = \|u\|'_d/d$ is seen to bound the rate of growth of the solutions. Indeed, the theorem in [10] states that if the sup-norm of the initial data is small, then solutions satisfy the bound

$$TV\big(u(\cdot,t)\big) \leq TV\big(u_0\big)\ \exp(k\rho t) + O(\epsilon),$$

where $k = 8\Lambda\lambda$ depends only on the flux. In this paper, we show that this bound is sharp; our solution satisfies

$$TV\big(u(\cdot,t)\big) \geq TV\big(u_0\big)\ \exp(\rho t),$$

so that the ratio again determines the rate of growth of solutions (we have $\Lambda = \lambda = 1$). Thus control of the d-norm leads to control of the time of existence of solutions.

When the coefficients are $\{1, 0, -1\}$, the behavior is very different. We may view this system as a completely linearly degenerate model for the equations of gas dynamics, as these are the coefficients for the Lagrange equations. In this case, we find that there are time-periodic solutions, as well as solutions which correspond to irrational rotations of the circle. These results indicate that there may be periodic solutions to the Euler equations, as well as other solutions which do not decay.

Although periodic solutions to 2×2 systems have been previously constructed [3, 4], those mechanisms for periodicity are different from the present case; indeed, those systems are genuinely nonlinear with trivial mode-mode interactions. In that case, the change in wave-speeds is exploited to piece together solutions consisting entirely of centered rarefaction waves.

These different types of solutions are not mutually exclusive; indeed, there are flux functions for which there are solutions which have exponential growth and decay, as well as periodic and irrationally oscillating solutions. The exact solutions can be described by a family of linear maps G_β of $\mathbf{R}^2$ parameterized by a number β, and with determinant 1. The dynamics of these maps are then determined by the eigenvalues, which may be real, corresponding to growth and decay, or complex, corresponding to rotations. As β varies, the map G_β may bifurcate, leading to the different phenomena we have described. Depending on the chosen flux, all of these phenomena may occur for one system. We remark that all of the initial data leading to these solutions may be chosen arbitrarily small in any periodic p-norm, showing sensitive dependence on initial data.

The paper is arranged as follows: in section 2, we recall the solution to the Riemann problem and the basic interaction estimate. In section 3, we construct the exact solutions which are to be analyzed. In section 4, we consider a specific system and find strongly unstable solutions. In section 5, we consider systems with a Riemann coordinateand describe in detail the exact solutions constructed above. Finally, in the appendix we show that the system used in section 4 exists; we were not able to write this system down explicitly.

2. Preliminaries. We briefly recall the construction of solutions to the Riemann problemand interactions of nonlinear waves. The Riemann problem is the problem obtained by taking Cauchy data consisting of two constant states, say u_L and u_R. It is solved by resolving the solution into constant states separated by N nonlinear waves, each wave corresponding to an eigenvalue of the flux matrix Df. These waves are either rarefactions or shocks when the field is genuinely nonlinear, $r^k \cdot \nabla \lambda_k > 0$, or contact discontinuities for a linearly degenerate field.

A k-rarefaction is a continuous expanding fan for which the state u_ϵ propagates along the characteristic given by $x/t = \lambda_k(u_\epsilon)$. Here the state u_ϵ lies on the integral curve of the eigenvector field $r^k(u)$ through the left state, that is, on the solution curve

$$\frac{du}{d\epsilon} = r^k(u), \quad \text{with} \quad u(0) = u_L.$$

This makes sense only if the wave-speed $\lambda_k(u_\epsilon)$ increases along this integral curve, which only accounts for positive values of the parameter ϵ. In order to pick up the other states to which u_L may be connected, we must consider discontinuous weak solutions, i.e., shocks. By considering the weak (integral) formulation of the conservation law, it is easy to see that discontinuities must satisfy the Rankine–Hugoniot relation

$$s(u_R - u_L) = f(u_R) - f(u_L),$$

where u_R and s are the right state and shock speed, respectively. This relation defines one curve in each family, tangent to the corresponding rarefaction curve. The entropy condition, which determines physically relevant shocks, selects that part of the curve along which λ_k decreases from left to right.

Combining the admissible parts of the shock and rarefaction curves, we complete the locus of states which can be connected to u_L by a (weak) nonlinear k-wave. It is well known that for a genuinely nonlinear field, the composite k-wave curve through a point is C^2 with discontinuous third derivative. The strength of a wave is given by the (signed) difference of a parameter which is increasing along the wave curve.

If the kth field is linearly degenerate, that is $r^k \cdot \nabla \lambda_k \equiv 0$ in a neighborhood, then the shock and rarefaction curves coincide, and the nonlinear wave is a jump discontinuity which propagates with speed $\lambda_k(u_L)$. In this case, the wave curve is exactly the integral curve of r^k and is as smooth as the flux. In particular, in our examples for which all eigenvalues are constant, and hence linearly degenerate, all nonlinear waves resolving Riemann problems are contacts which propagate with the constant speed λ_k. There are thus no nonlinear effects due to the expansion and compression of waves; rather, *all nonlinear effects come from interactions of waves of different families.*

Once we have constructed the wave curves, the general Riemann problem is solved by centering N nonlinear waves, one for each family, at the origin, and finding the correct intermediate states so that the extreme left and right states are those of the Riemann data. Note that we are using strict hyperbolicity here: since all waves are centered at the origin, these must appear in increasing order of wave-speed from left to right. This gives us a natural ordering of wave curves which must be taken into account when resolving wave interactions. Also, because the eigenvectors form a basis, the implicit function theorem yields a unique decomposition into waves and constant states [7].

The Riemann interaction problem is formulated as follows: suppose that the Riemann problems with data $\langle u_L, u_M\rangle$ and $\langle u_M, u_R\rangle$ are resolved into waves of known strength $\{\alpha_i\}$ and $\{\beta_i\}$, respectively. If the resulting Riemann problem with data $\langle u_L, u_R\rangle$ is resolved into waves $\{\epsilon_i\}$, the interaction problem is to estimate the wave strengths ϵ_i in terms of α_i and β_i. This problem has a satisfactory answer for weak waves, namely

$$\epsilon_i = \alpha_i + \beta_i + \sum_{j>k} \Lambda_i^{jk} \alpha_j \beta_k + O(SD). \tag{2.1}$$

Here the coefficients Λ_i^{jk} are the interaction coefficients determined by the Lie bracket, $[r^j, r^k] = \sum \Lambda_i^{jk} r^i$, while S is equivalent to the L^∞-norm of the solution and D is a quadratic error term measuring the error due to interactions [1, 12]. For our purposes, it is enough to note that $D = O(S^2)$, where S can be taken to be the maximum strength of interacting waves.

To interpret the estimate geometrically, we observe that the leading order effect of the interaction of a pair of waves from different families is simply the transposition of these waves, which is achieved by integrating along the wave curves in reversed order. The quadratic correction due to this reversal is then given by the Lie bracket of the vector fields, which yields the quadratic term in the estimate. We get the interaction estimate by including one bracket for each pair of transposed waves, namely α_j and β_k, where $j > k$, and separating these corrections into wave components. We remark that this interaction effect is determined by the geometry of the eigenvector fields, and is not related to the genuine nonlinearity of any individual wave family.

The interaction problem cannot in general be interpreted (locally) as the solution to a Riemann problem, except when all the pairwise interactions occur simultaneously. This can happen whenever a number of waves converge at a single point of space-time. In this case, a similar interaction estimate holds, and the continuation of the solution beyond the point of interaction is indeed an exact solution to the conservation law. In our examples, all waves are jump discontinuities, so we shall consider three waves in decreasing families converging to a single point, as in Figure 2.1, and refer to this as a triple interaction. Suppose the incident waves are γ, β, and α and these are, respectively, 3-, 2-, and 1-waves. After interaction, we resolve the solution into waves in increasing order of wave-speed. If waves are labeled as in the figure, then the outgoing wave strengths are given by

$$\begin{aligned} \alpha' &= \alpha + \Lambda_1^{32} \gamma\beta + O(S^3), \\ \beta' &= \beta + \Lambda_2^{31} \gamma\alpha + O(S^3), \quad \text{and} \\ \gamma' &= \gamma + \Lambda_3^{21} \beta\alpha + O(S^3), \end{aligned} \tag{2.2}$$

where $S = \max\{\alpha, \beta, \gamma\}$. Here we have normalized the eigenvectors so that all other interaction coefficients vanish at a point (see [11]). Figure 2.1 illustrates the interaction in two spaces, first by tracing the paths of the contacts ("characteristics") in space-time, and also by showing the constant states lying on the integral curves of eigenvectors in state space. Note that the characteristics are exactly as for a linear system, but the intermediate states are changing nonlinearly.

We shall use these triple interactions to build up periodic resonant solutions to our degenerate conservation laws, in which all interactions are of this type. These solutions will be piecewise constant, and we shall analyze them in some detail.

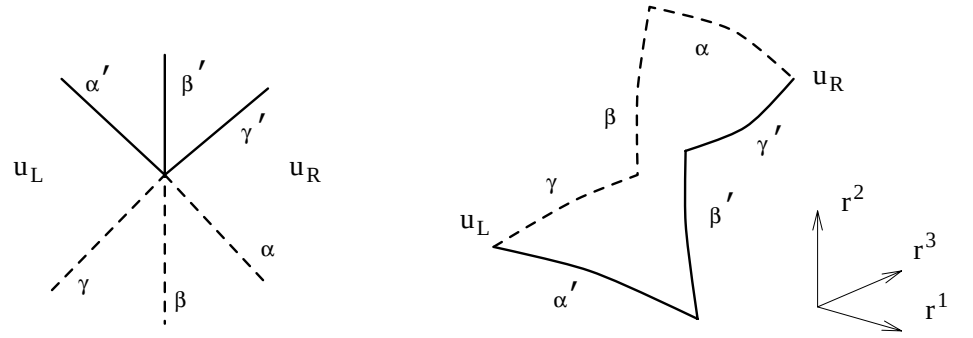


FIG. 2.1. *A triple interaction.*

3. Exact periodic solutions. We now construct exact solutions for certain Cauchy data for our degenerate conservation laws. Recall that these systems have constant wave-speeds, and the nonlinearity comes from the geometry of the eigenvector fields. We shall construct periodic solutions, although it is clear that these can be restricted to compact support. Our building blocks are the triple interactions constructed above.

In order to generate a repeating resonance pattern, we require that the ratio of differences of wave-speeds, $\frac{\lambda_3-\lambda_2}{\lambda_2-\lambda_1}$, be rational [5], and we shall take this to be unity. For simplicity, we assume that the eigenvalues of the conservation law are identically -1, 0, and 1. Thus all waves are contact discontinuities and propagate along straight line "characteristics" of slope 0 or ± 1 in space-time, and we set $\Delta t = \Delta x$. This is analogous to solutions of linear systems, for which singularities propagate along characteristics, $x = 0$ or $x = \pm t$, respectively. In our case, the singularities are jump discontinuities which, although they propagate at fixed speed, change strength after passing through each interaction.

We now set up a periodic resonant wave pattern. This is a local construction and all waves will be assumed to be weak, so that the solution is restricted to a neighborhood of state space, and Riemann problems can be solved. We start with a localized configuration of waves and extend this by periodicity. Referring to Figure 3.1, we choose a reference state, which we label 1. We then find state 2 on the 2-wave curve of 1, so that this 2-wave has strength δ_1, say. Choose another state 5, and connect this to state 6 by another 2-wave, say of strength δ_2. We now let Δx be a small positive number: our Cauchy data will be constant on intervals of length Δx. Indeed, we define the Cauchy data to take on the states 1, 2, 5, and 6 on consecutive intervals of length Δx, respectively, and extend this data periodically. We shall choose the states 1 and 5 and strengths δ_1 and δ_2 to obtain different types of behavior.

Once the Cauchy data has been prescribed, we can determine the solution as follows. Up to time $t = \Delta x$, we simply resolve the Riemann problems into constant states separated by waves, which by degeneracy are all contacts. By our choice of states, the Riemann problems $\langle 1, 2\rangle$ and $\langle 5, 6\rangle$ are resolved into a single 2-wave. Resolving the Riemann problems $\langle 6, 1\rangle$ and $\langle 2, 5\rangle$, we get three waves emanating from each point $x = 2k\Delta x$. At time $t = \Delta x$ triple interactions occur, and the solution is of the same form as the Cauchy data. Continuing in this way, a resonant periodic wave interaction pattern is formed. Figure 3.1 illustrates the resonance pattern by showing the characteristics and a projection of state space, where we integrate along the wave curves. We have labeled the constant states to show how they change, and

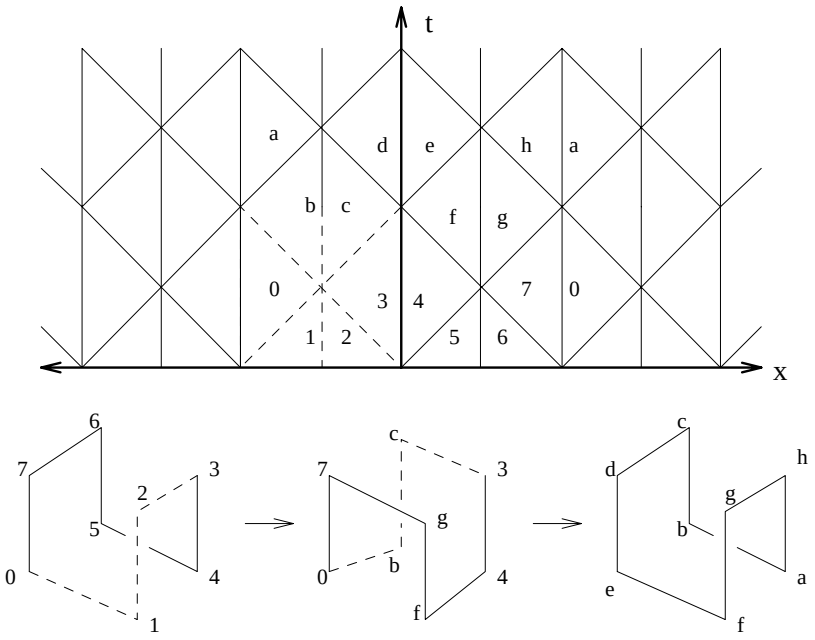


FIG. 3.1. *A resonant interaction pattern.*

the dashed lines show a single interaction.

It is clear that all interactions of waves are triple interactions, occurring exactly at the points $(x,t) = (k\Delta x, j\Delta x)$, for $j+k$ even. Moreover, the solution at each interaction time $t = j\Delta x$ is of the same form as the Cauchy data, and the wave pattern is repeated, although the intermediate states and wave strengths themselves change in time. This is to be contrasted with the usual behavior for nonlinear conservation laws, in which the wave patterns change and may become very complicated due to the expansion of rarefactions and formation of shock waves. We emphasize that we have constructed an exact solution to the degenerate conservation law and not an approximation.

At each fixed time, our solution is space-periodic with period $4\Delta x$ and takes on 8 (or 4) constant values. If these intermediate values are known, it is easy to write down the L^p-norm of the solution (as a $4\Delta x$-periodic function.) For example, labeling the states as in Figure 3.1, at time $t = 2\Delta x$ we have

$$\|u\|_p^p = |b|^p + |c|^p + |f|^p + |g|^p.$$

Similarly, by knowing the wave strengths at any time, we can estimate the L^p-norm of the solution. In fact, since the wave pattern and period of the solution do not change, a knowledge of the minimum and maximum wave strengths (the sup-norm) will allow us to deduce estimates for the other L^p-norms as well.

Temple and Young have introduced a norm, called the d-norm, which measures the local variation of the solution [10]. This norm is defined for fixed (small) d as

$$\|u\|_d' = \sup_a TV_{[a,a+d]}(u(\cdot)) \tag{3.1}$$

to be the maximum variation of a function $u(x)$ over intervals of length d. Clearly, if we set $d = 4\Delta x$, then for our examples the d-norm is just the total variation per

period of the solution. In [10], it was shown that the size of d and the corresponding value of the d-norm are critical in determining the growth of solutions.

Thus far we have set up the wave pattern without specifying the interaction coefficients or the exact Cauchy data. We shall choose the states 1 and 5 and wave strengths δ_1 and δ_2 to give different effects depending on the values of the interaction coefficients. We are primarily interested in two cases: all three interaction coefficients are positive and one coefficient vanishes.

Before proceeding with specific choices, we recall the interaction estimate (2.3):

$$\begin{aligned} \alpha' &= \alpha + \Lambda_1 \gamma\beta + O(S^3), \\ \beta' &= \beta + \Lambda_2 \gamma\alpha + O(S^3), \quad \text{and} \\ \gamma' &= \gamma + \Lambda_3 \beta\alpha + O(S^3), \end{aligned} \tag{3.2}$$

where α, β, and γ are 1-, 2-, and 3-waves, respectively, and $'$ refers to the outgoing wave, as in Figure 2.1. Assuming that $\Lambda_i \geq 0$ and ignoring cubic effects, we see that growth can occur in each family with a nonzero coefficient if the product $\alpha\beta\gamma$ is positive. That is, each outgoing wave will be of the same sign but have greater strength than the corresponding incident wave. We shall say that magnification occurs in the interaction. This product is positive if all three incident waves are positive, or if exactly two are negative. This gives us room to set up the resonance pattern so that each triple interaction yields magnification of wave strengths. Similarly, if the product $\alpha\beta\gamma$ is negative, then after interaction each wave is weaker but of the same sign, and we find resonant solutions which decay.

A careful check using Figure 3.1 reveals that if consecutive waves in each family alternate in sign, then the sign of the product of incident waves is the same at each lattice point. For example, with the notation of Figure 3.1, we take the 1-waves 23 and 67 to have opposite sign, and similarly the 2-waves 12, 34, 56, and 70 to alternate in sign, etc. Here 23 refers to the 1-wave between constant states 2 and 3, and so on. If this product of incident waves is positive we expect growth, and if it is negative we expect decay. In the following sections we shall examine the behavior of solutions for some specific values for the interaction coefficients.

4. Instability of solutions. We now make specific assumptions on the interaction coefficients and study the consequences of these assumptions. In this section we shall suppose that the interaction coefficients are all positive. The existence of a flux function which has constant wave-speeds and these coefficients is established in the appendix. The system can thus be described as follows: the wave-speeds are given by -1, 0, and 1, and the interaction coefficients Λ_i are positive. The following theorem holds for this system.

THEOREM 1. *There are solutions with initial data of bounded variation and arbitrarily small L^p-norm which grow arbitrarily in any finite time. More precisely, given positive constants τ, ϵ, and $K \gg 1$, there are some $t_e < \tau$ and weak solutions $u(x,t)$ defined for $t \in [0, t_e)$ satisfying*

$$\begin{gathered} \|u(\cdot,0)\|_\infty < \epsilon, \quad \|u(\cdot,0)\|_1 < \epsilon, \quad \textit{and} \\ 8/\Lambda_m \leq TV\big(u(\cdot,0)\big) \leq 10/\Lambda_m, \end{gathered}$$

where $0 < \Lambda_m < \min \Lambda_i$, while for some $t_b < t_e$, we have

$$\begin{aligned} \|u(\cdot,t_b)\|_\infty &> K \|u(\cdot,0)\|_\infty, \\ \|u(\cdot,t_b)\|_1 &> K \|u(\cdot,0)\|_1, \quad \textit{and} \\ TV\big(u(\cdot,t_b)\big) &> K\, TV\big(u(\cdot,0)\big). \end{aligned}$$

Similar statements hold for all other L^p-norms.

We note that although these solutions are highly unstable, they do not necessarily become infinite in finite time. That is, although the amplitudes can be arbitrarily magnified, our construction breaks down before solutions become infinite. Indeed, once the amplitude becomes large, the global structure of the wave curves becomes important and the Riemann problem is not necessarily well-posed. This is the same type of behavior seen in the constructions of geometric optics [5, 6].

It is remarkable that instability occurs for data which has total variation larger than $8/\Lambda_m$, when contrasted with the result of the author that solutions are stable as long as the initial total variation is less than $1/3\Lambda_M$, where $\Lambda_M = \max \Lambda_i$, and the sup-norm is small enough [12]. In this case, although a resonating pattern can be established and rapid growth may occur, there are not enough waves initially to preserve the resonance, so that the different families will separate, halting the growth process. In general, once the families separate, genuine nonlinearity takes over and solutions decay [8].

This behavior is clarified by the growth rate of solutions established in [10] and its relation to the d-norm defined there. The d-norm, given by (3.1), measures the "local" variation of the solution and provides a scale for the short-range effects of interactions, which do not cause growth in the solution. However, when the total variation is large, the long-range effects of interactions (namely those accumulating over multiple d-intervals) may dominate and cause instability. In our examples, the growth rate is a constant multiple of the ratio $\|u_0\|'_d/d$, which in turn determines the time of existence without explicit reference to the total variation and sup-norm. In particular, if we let $d \to 0$ faster than $\|u_0\|'_d$, then our solutions explode in arbitrarily short time. The following corollaries are evident from the proof of the theorem.

COROLLARY 1. *The d-norm determines the rate of growth and time of existence of solutions. That is, the amplitude $\eta(t)$ of the solution satisfies*

$$\eta(t) \geq \frac{\eta(0)}{1 - k\,\rho\,t},$$

where ρ is the ratio $\|u\|'_d/d$ and k is a constant depending only on the flux.

COROLLARY 2. *There are space-periodic solutions with arbitrarily small L^p-norms (measured over one period) which grow arbitrarily in finite time. That is, there are periodic solutions $u(x,t)$ satisfying*

$$\|u(\cdot,0)\|_p < \epsilon, \quad \textit{while} \quad \|u(\cdot,t_b)\|_p > K\,\|u(\cdot,0)\|_p$$

for each $1 \leq p \leq \infty$.

COROLLARY 3. *There are periodic solutions (to the same conservation law) defined for all time, which decay at the rate $1/(1+Kt)$, for any given constant K.*

Proof. We shall use the exact solutions constructed above and analyze changes in wave strengths after carefully choosing the initial data. In the notation of Figure 2.1, recall that we have the freedom to choose constant states which we labeled 1 and 5, and initial 2-wave strengths δ_1 and δ_2. We shall make these choices so that all waves have approximately the same strength, while waves in each family have alternating signs. Let $\eta > 0$ denote a small number to be determined, and fix state 1 inside a neighborhood of the origin. Let $\delta_1 = \delta_2 = \eta$, which determines states 2 and 6, once 5 has been chosen. We choose state 5 by setting the strengths of the waves 23, 34, and 45 to be η, $-\eta$ and $-\eta$, respectively. This fixes our initial data and implies that the waves 67, 70, and 01 have *approximate* strength $-\eta$, $-\eta$, and η, respectively. We now

find $\eta_0 \approx \eta$ so that all initial wave strengths are between η_0 and $5\eta_0/4$, where η_0 will be chosen later.

With these choices, we see that the product of strengths of waves entering each interaction of the lattice is positive, so that magnification occurs at each interaction. In order to see that unstable growth occurs, we now define η_q to be the minimum strength of all waves at time $t = 2q\,\Delta x+$, that is, after each wave has passed through at least q interactions. Then, assuming that the waves remain weak, for some positive $\Lambda' < \Lambda_m < \Lambda_i$, we have

$$\eta_{q+1} \geq \eta_q + \Lambda_m\,\eta_q^2 \geq \eta_q + \Lambda'\,\eta_q\,\eta_{q+1} \tag{4.1}$$

for each q. This is true by the interaction estimate (2.3), where we observe that the error $O(SD)$ is cubic in η_q, so that it is dominated by the quadratic part. Now, by comparison with the exact solution of the corresponding difference equation, we get

$$\eta_n \geq \frac{\eta_0}{1 - \Lambda'\,\eta_0\,n} \quad \text{for each } n. \tag{4.2}$$

We obtain solutions with compact support simply by cutting off the wave pattern for some $|x| > L$. Note that if we do this, resonance occurs in a region containing the set $\mathcal{S} = \{(x,t) | -L+t < x < L-t\}$. Our calculation of η_n remains valid as long as η_n measures the strength of waves contained in $\mathcal{S}$. We now estimate the norms for the solution by waves contained in $\mathcal{S}$. The sup-norm is measured by the maximum wave strength, and the total variation is measured by the sum of (absolute) wave strengths. We thus have

$$\begin{aligned} \eta_0 \leq \|u(.,0)\|_\infty \leq 5\eta_0/4 \quad &\text{and} \\ 4L\,\eta_0/\Delta x \leq TV\big(u(.,0)\big) \leq 5L\,\eta_0/\Delta x, & \end{aligned}$$

as there are four waves per interval of length $2\Delta x$. The L^1-norm is given by

$$\|u(.,0)\|_1 = 2L/4\Delta x \cdot \int_0^{4\Delta x} |u| \;\sim \eta_0 L.$$

Similarly, the norms at time $t < L$ satisfy

$$\begin{aligned} \|u(.,t)\|_\infty &\geq \eta_n, \\ TV\big(u(.,t)\big) &\geq 4(L-t)\,\eta_n/\Delta x, \quad \text{and} \\ \|u(.,t)\|_1 &\geq (L-t)\,\eta_n, \end{aligned}$$

where n is given by $n = [t/2\Delta x]$. Now, given any constant K, instability in all norms will follow as long as $\eta_n > \eta_0 \cdot 5KL/4(L-t)$, or by (4.2), if

$$\frac{1}{1 - \Lambda'\,\eta_0\,n} > \frac{5\,K\,L}{4\,(L-\tau)}, \tag{4.3}$$

while also $t < \tau < L$. Therefore we take $L > \tau$ and choose η_0 and Δx such that the relations

$$\Lambda'\,\eta_0\,n \to 1 \quad \text{and} \quad t \to \tau$$

are equivalent. This can be accomplished by defining $\rho = 2/\Lambda'\,\tau$ and setting $\eta_0 = \rho\,\Delta x$. We now choose Δx so small that (4.3) holds, while η_{n+2} remains small enough

that all estimates used above still hold. It is clear that the sup- and L^1-norms of the initial data can be made arbitrarily small, while the initial total variation is bounded below by $4L\rho > 8/\Lambda'$. Also, since the waves remain a fixed distance Δx apart while they grow in size, we have instability of the solution in all L^p-norms. Similar estimates hold for space-periodic solutions.

We can interpret our choice of the parameter ρ in terms of the d-norm as follows: with $d = 4\Delta x$ set to one period, we see that the d-norm satisfies

$$8\eta_0 \leq \|u(.,0)\|'_d \leq 10\eta_0$$

or equivalently the ratio

$$2\rho \leq \frac{\|u(.,0)\|'_d}{d} \leq 2.5\rho.$$

That is, our choice of ρ is just our choice of the ratio $\|u\|'_d/d$, which in turn determines the growth rate of solutions. Indeed, we can rewrite (4.2) as

$$\eta_n \geq \frac{\eta_0}{1 - \Lambda'\rho t/2},$$

and Λ' is a constant depending only on the flux.

The proof of the third corollary proceeds as above, except that we change the sign of the waves in one of the families in the initial data, say setting 45 to η. This makes all triple interactions lead to decay, so that the estimates will be reversed. Indeed, if χ_i denotes the maximum absolute wave strength, then we will get $\chi_{q+1} \leq \chi_q - \bar{\Lambda}\chi_{q+1}\chi_q$ for $\bar{\Lambda} \geq \Lambda_M$, so that

$$\chi_n \geq \frac{\chi_0}{1 + \bar{\Lambda}\,\chi_0\,n}.$$

We can now scale the initial data by taking $\chi_0 = \rho\Delta x$ with $\rho = 2K/\bar{\Lambda}$. We omit the details. □

The growth and decay of solutions is analogous to that of solutions of the equation $y' = y^2$ depending on the sign of the initial data. The extra freedom in choosing the rate of growth or decay is a consequence of total linear degeneracy: we are able to move the contacts closer together so that oscillations are more rapid without changing the basic wave pattern. In a genuinely nonlinear system, increasing the rate of oscillations in a fixed interval just leads to faster decay in each family as shock and rarefactions collide at a higher rate. We control the growth rate by choosing the d-norm, which in turn measures the scaling $\eta_0 \sim \Delta x$.

5. Time-periodic solutions. In realistic systems, there are usually extra constraints placed on the system by physical principles. One of the most common of these is to assume the existence of a Riemann coordinate. For these systems, solutions are more stable than the examples given above, and Temple and Young have recently shown that growth of solutions to 3×3 systems with a single Riemann coordinate can be at most exponential [10]. In this section we show that this theorem is sharp in the absence of other assumptions. The variety of qualitative phenomena present even when there is a Riemann coordinate suggests that further assumptions need to be made in order to determine specific properties of solutions.

A Riemann coordinate for the kth field is a function ρ whose gradient is orthogonal to all other eigenvectors: $r^j \cdot \nabla\rho = 0$ for each $j \neq k$. The existence of a Riemann

coordinate implies that the corresponding wave family decouples from the system and is equivalent to the vanishing of one of the interaction coefficients. The definition implies that $\nabla\rho$ is a kth left eigenvector, which in turn allows for a trivial row of the matrix $A = Df$ in some coordinate system, corresponding to a dimension reduction.

We shall consider three systems, with different flux coefficients, solutions of which exhibit different qualitative behavior. For definiteness we suppose that we have a 2-Riemann coordinate. The flux functions and their respective triples of interaction coefficients are given by

$$f_1(u,v,w) = \begin{pmatrix} we^{2v} \\ 0 \\ ue^{-2v} \end{pmatrix} \quad \text{with coefficients} \quad \{1,0,-1\}, \tag{5.1}$$

$$f_2(u,v,w) = \begin{pmatrix} w+2uv \\ 0 \\ u(1-4v^2)-2vw \end{pmatrix} \quad \text{with} \quad \{1,0,1\}, \quad \text{and} \tag{5.2}$$

$$f_3(u,v,w) = \begin{pmatrix} uv+we^{v} \\ 0 \\ u(1-v^2)e^{-v}-vw \end{pmatrix} \quad \text{with} \quad \left\{1-\frac{v}{2}, 0, \frac{-v}{2}\right\}. \tag{5.3}$$

We have the following theorem.

THEOREM 2. *For the system with flux f_1, there are space-periodic solutions of period π_x which are periodic in time with period $\pi_t = n\pi_x$ for integers n. These solutions are not stable under perturbations of the initial data, and we also have oscillating solutions corresponding to rotation by an irrational angle.*

For the system with flux f_2, there are periodic solutions which grow exponentially in those families not associated to the Riemann coordinate. Similarly, there are also solutions which decay exponentially in these families.

For the system with flux f_3, there are solutions exhibiting all of the above behaviors. That is, depending on the initial data, the solution may be exponentially growing or decaying, be periodic, or may oscillate as an irrational rotation.

These phenomena are local, in that the initial data may be arbitrarily small in any (periodic) L^p-norm. The growth of decay can be measured in these L^p-norms, while the oscillating solutions remain on an ellipse.

We remark that as in the previous examples, the linear degeneracy gives us the freedom to choose the rates at which these solutions decay or grow in time. We expect that if one weakens the hypothesis of degeneracy, it is still possible to get exponential growth of the solution, albeit not with an arbitrary rate. The variety of behaviors obtained for these simple examples again demonstrates the necessity of knowing exactly what the interaction coefficients are in order to make general statements about properties of solutions to conservation laws. We shall see that because of the simplicity of our examples, we do not require the solutions to be small, and indeed the solutions are globally defined.

Our analysis of system (5.2) shows that the result of [10] on stability of solutions is sharp: that result says that for data with small sup-norm, the total variation grows at most exponentially:

$$TV\big(\mathbf{u}(\cdot,t)\big) \le TV\big(\mathbf{u}_0\big)\ \exp(k\delta t/d) + O(\epsilon),$$

where $k = 8\Lambda\lambda$ (Λ and λ are the interaction coefficient and CFL number); δ is the d-norm of the Riemann coordinate; and ϵ is the amplitude of the initial data. Note that for (5.2), we have $\Lambda = \lambda = 1$.

COROLLARY 4. *The d-norm determines the growth rate of solutions to* 3×3 *systems with one Riemann coordinate. That is, there is a solution of* (5.2) *whose total variation satisfies*

$$TV\big(u(\cdot,t)\big) \geq TV\big(u_0\big)\ \exp(\delta t/d),$$

where $\delta = \|v_0\|'_d$ *is the d-norm of the Riemann coordinate* v*. A similiar statement holds for* w*. Here the vector* $\mathbf{u}_0 = (u_0, v_0, w_0)$ *is the initial data, and this can be taken to have arbitrarily small sup-norm.*

Proof. The proof is an exact calculation of the wave strengths for the exact periodic solutions constructed earlier. We shall consider a general case, substituting in the specific fluxes at the end.

By inspection, we see that the matrix

$$A = \begin{pmatrix} \sigma & C_1 & \tau \\ 0 & \nu & 0 \\ \frac{1-\sigma^2}{\tau} & C_2 & -\sigma \end{pmatrix}$$

has eigenvalues -1, ν, and 1, and left eigenvector $(0,1,0)$ corresponding to ν, which we take to be zero. With σ and τ yet to be chosen, we find C_1 and C_2 so that the rows become gradients. We compute the eigenvectors directly and the interaction coefficients according to the definition $\Lambda_i^{jk} = l_i \cdot [r^j, r^k]$. We then choose the functions σ and τ so that these take on the desired values.

It suffices to take σ and τ to be functions of v alone and choose C_1 and C_2 so that the rows of A are curl-free. For convenience we define the 2×2 matrix

$$\hat{A}(v) = \begin{pmatrix} \sigma(v) & \tau(v) \\ \frac{1-\sigma(v)^2}{\tau(v)} & -\sigma(v) \end{pmatrix},$$

and we take

$$\begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \sigma & \tau \\ \frac{1-\sigma^2}{\tau} & -\sigma \end{pmatrix}' \begin{pmatrix} u \\ w \end{pmatrix} = \hat{A}'(v) \begin{pmatrix} u \\ w \end{pmatrix}. \tag{5.4}$$

The full eigensystem of A is given by

$$\begin{array}{lccc} \lambda & -1 & 0 & 1 \\ r & \begin{pmatrix} \tau \\ 0 \\ -1-\sigma \end{pmatrix} & \begin{pmatrix} a \\ 1 \\ b \end{pmatrix} & \begin{pmatrix} \tau \\ 0 \\ 1-\sigma \end{pmatrix} \\ l & (\frac{1-\sigma}{2\tau} \ .. \ \frac{-1}{2}) & (0\ 1\ 0) & (\frac{\sigma+1}{2\tau} \ .. \ \frac{1}{2}) \end{array},$$

where a and b satisfy

$$\hat{A} \begin{pmatrix} a \\ b \end{pmatrix} + \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = 0, \tag{5.5}$$

and so, since $\hat{A}^{-1} = \hat{A}$ and by (5.4),

$$\begin{pmatrix} a \\ b \end{pmatrix} = -\hat{A}^{-1} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = -\hat{A}(v)\ \hat{A}'(v) \begin{pmatrix} u \\ w \end{pmatrix}. \tag{5.6}$$

For this class of systems, we can integrate the wave curves and describe Riemann solutions and interactions explicitly, and thus find exact algebraic expressions for the solutions constructed in section 3. Since σ and τ are functions of v only, and this is the 2-Riemann coordinate, the 1- and 3-wave curves through $\{u_0, v_0, w_0\}$ are simply straight lines. If we define

$$\hat{r}^+(v) = \begin{pmatrix} \tau(v) \\ 1-\sigma(v) \end{pmatrix} \quad \text{and} \quad \hat{r}^-(v) = \begin{pmatrix} \tau(v) \\ -1-\sigma(v) \end{pmatrix}$$

to be the eigenvectors of $\hat{A}$ corresponding to the eigenvalues ± 1, then we can describe the 1- and 3-wave curves by

$$\begin{pmatrix} u \\ w \end{pmatrix} = \begin{pmatrix} u_0 \\ w_0 \end{pmatrix} + \epsilon \hat{r}^{\pm}(v_0) \quad \text{and} \quad v = v_0,$$

respectively, where ϵ is the wave strength. The 2-wave curve is found by noticing that we can use v as dependent variable and rewriting the system $u' = a$, $w' = b$. According to (5.4) and (5.5), the 2-wave curve satisfies

$$\frac{d}{dv}\left[\hat{A}(v)\begin{pmatrix} u \\ w \end{pmatrix}\right] = 0,$$

so the wave curve is given by

$$\hat{A}(v)\begin{pmatrix} u \\ w \end{pmatrix} = \hat{A}(v_0)\begin{pmatrix} u_0 \\ w_0 \end{pmatrix}$$

with wave strength $\epsilon = v - v_0$.

We can now write the complete solution to the Riemann problem: if states $\{u_L, v_L, w_L\}$ and $\{u_R, v_R, w_R\}$ are connected by 1-, 2-, and 3-waves α', β', and γ' respectively, then the wave strengths are determined by

$$\hat{A}(v_L)\left\{\begin{pmatrix} u_L \\ w_L \end{pmatrix} + \alpha' \hat{r}^-(v_L)\right\} = \hat{A}(v_R)\left\{\begin{pmatrix} u_R \\ w_R \end{pmatrix} - \gamma' \hat{r}^+(v_R)\right\} \tag{5.7}$$

and $\beta' = v_R - v_L$. The same expression can be used to find the right state once we are given a left state and wave strengths. Similarly, if $\{u_L, v_L, w_L\}$ and $\{u_R, v_R, w_R\}$ are connected by 3-, 2-, and 1-waves γ, β, and α, respectively, and converging at a point as in Figure 2.1, we get the expression

$$\hat{A}(v_L)\left\{\begin{pmatrix} u_L \\ w_L \end{pmatrix} + \gamma \hat{r}^+(v_L)\right\} = \hat{A}(v_R)\left\{\begin{pmatrix} u_R \\ w_R \end{pmatrix} - \alpha \hat{r}^-(v_R)\right\}.$$

Combining these expressions, we can describe the triple interaction depicted in Figure 2.1 by solving

$$\alpha' \hat{r}^-(v_L) - \gamma' \hat{r}^+(v_R) = \alpha \hat{r}^-(v_R) - \gamma \hat{r}^+(v_L) \quad \text{and} \quad \beta' = \beta, \tag{5.8}$$

where $v_R = v_L + \beta$. We stress that this is an exact formula, and it is also global; that is, it holds for incident waves of arbitrary strength.

We can compute the interaction coefficients directly: clearly $r^0 \cdot \nabla r^{\pm} = r^{\pm\prime}$, and from (5.6) we get

$$\hat{r}^{\pm} \cdot \nabla_{\{u,w\}} \begin{pmatrix} a \\ b \end{pmatrix} = -\hat{A}\, \hat{A}'\, \hat{r}^{\pm}.$$

Using these facts we get

$$\Lambda_1 = \frac{\tau\sigma' - (\sigma-1)\tau'}{2\tau}, \quad \Lambda_2 = 0 \quad \text{and} \quad \Lambda_3 = \frac{\tau\sigma' - (\sigma+1)\tau'}{2\tau}.$$

If we now choose $\sigma(v) = 0$ and $\tau(v) = e^{2v}$, we get flux f_1, while if $\tau(v) = 1$ and $\sigma(v) = 2v$, we get f_2. Flux f_3 is obtained by taking $\sigma(v) = v$ and $\tau(v) = e^v$. The corresponding interaction coefficients are then as in (5.1)–(5.3).

We now construct periodic initial data for the resonant pattern used previously and examine the behavior of these solutions. We shall consider only the simplest cases, but because the calculations are exact, it should be possible to treat more general cases.

Referring to Figure 3.1 and using the notation described earlier, we choose our initial data as follows. Fix state 1 and numbers α, β, and γ, which will refer to 1-, 2-, and 3-waves, respectively. As before, we choose the wave strengths 12, 23, 34, 45, and 56, which determine states 2 through 6, and resolve the Riemann problem $\langle 6, 1\rangle$ to determine the remaining wave strengths. If we define strengths 12, 23, 34, 45, and 56 to be β, α, $-\beta$, $-\gamma$, and β, respectively, then by a direct calculation using (5.7), we see that strengths 67, 70, and 01 are given *exactly* by $-\alpha$, $-\beta$, and γ, respectively. Moreover, we have that $v_4 = v_5 = v_0 = v_1$ and $v_2 = v_3 = v_6 = v_7 = v_1 + \beta$, where v_i is the v-component of state i, and these values do not change in time. This is to be expected from the "decoupled" equation $v_t = 0$ corresponding to the Riemann coordinate.

As a consequence of these observations, we see that we can describe the solution fully for all time with a knowledge of v_1 and v_2 and the strengths of the 1- and 3-waves in the solution at each time. In other words, once state 1 and strength β have been chosen, all we need to describe the system fully is a knowledge of the 1- and 3-wave strengths (corresponding to α and β) at each time. We shall use (5.8) to describe these changing wave strengths and choose different values for α, β, and γ to find solutions with different properties.

To this end, we let α_k denote the strength of the 1-wave in the solution (with the same sign as $\alpha = \alpha_0$) for time $k\Delta x < t < (k+1)\Delta x$ and define γ_k similarly. Then a representative triple interaction at time $(2k+1)\Delta x$ has waves γ_{2k}, β, and α_{2k} entering and α_{2k+1}, β, and γ_{2k+1} leaving, so that according to (5.8), we have

$$\alpha_{2k+1}\hat{r}^-(v_1) - \gamma_{2k+1}\hat{r}^+(v_2) = \alpha_{2k}\hat{r}^-(v_2) - \gamma_{2k}\hat{r}^+(v_1).$$

On the other hand, an interaction at time $2k\Delta x$ has waves $\pm\gamma_{2k+1}$, $-\beta$, and $\mp\alpha_{2k+1}$ entering, so that in this case, we get

$$\alpha_{2k+2}\hat{r}^-(v_2) + \gamma_{2k+2}\hat{r}^+(v_1) = \alpha_{2k+1}\hat{r}^-(v_1) + \gamma_{2k+1}\hat{r}^+(v_2).$$

We express the foregoing in matrix form as follows. We define

$$M = \begin{pmatrix} \hat{r}_1^- & \hat{r}_2^+ \end{pmatrix}, \quad N = \begin{pmatrix} \hat{r}_2^- & \hat{r}_1^+ \end{pmatrix}, \quad \text{and} \quad S = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix},$$

where the vectors $\hat{r}_i = \hat{r}(v_i)$ are viewed as columns. We then rewrite the above as

$$M\,S \begin{pmatrix} \alpha_{2k+1} \\ \gamma_{2k+1} \end{pmatrix} = N\,S \begin{pmatrix} \alpha_{2k} \\ \gamma_{2k} \end{pmatrix} \quad \text{and} \quad N \begin{pmatrix} \alpha_{2k+2} \\ \gamma_{2k+2} \end{pmatrix} = M \begin{pmatrix} \alpha_{2k+1} \\ \gamma_{2k+1} \end{pmatrix}.$$

This in turn yields

$$\begin{pmatrix} \alpha_{2k+2} \\ \gamma_{2k+2} \end{pmatrix} = G(v_1, v_2) \begin{pmatrix} \alpha_{2k} \\ \gamma_{2k} \end{pmatrix},$$

where the matrix $G(v_1, v_2)$ is given by

$$G(v_1, v_2) = N^{-1}\, M\, S\, M^{-1}\, N\, S. \tag{5.9}$$

The wave strengths α_{2k} and γ_{2k} are now easily calculated, namely

$$\begin{pmatrix} \alpha_{2k} \\ \gamma_{2k} \end{pmatrix} = G(v_1, v_2)^k \begin{pmatrix} \alpha \\ \gamma \end{pmatrix},$$

and α_{2k+1} and γ_{2k+1} are found similarly.

In order to complete the description of our solution to the conservation law, we calculate the powers of the matrix G. By inspection, G has determinant 1, so we can find the eigenvalues of G by a knowledge of its trace. Indeed, if $tr(G) = 2\mu$, then the eigenvalues are $\mu \pm \sqrt{\mu^2 - 1}$. Thus if $|\mu| < 1$, these are complex (with modulus 1), corresponding to oscillations, while if $|\mu| > 1$, they are real, corresponding to growth and decay.

As we have seen, the 2-wave strength β does not change with time, and the behavior of the solution to the conservation law is given by knowledge of the wave strengths α_k and γ_k. This can be regarded as a discrete dynamical system determined by the 2-parameter family of linear maps $G(v_1, v_2)$. With this point of view, we can study the stability and bifurcation properties of G. This in turn leads to statements about the stability or lack thereof for solutions to the conservation law.

We now restrict ourselves to the fluxes specified earlier, and describe the solutions. First, for the flux f_1, we had $\sigma(v) = 0$ and $\tau(v) = e^{2v}$, and we set $v_2 = v_1 + \beta$, so that $\tau(v_2) = e^{2\beta}\tau(v_1)$. After substituting and manipulating, we get

$$G(v_1, v_2) = \begin{pmatrix} \mathrm{sech}\beta & e^{-\beta}\tanh\beta \\ -e^{\beta}\tanh\beta & \mathrm{sech}\beta \end{pmatrix}^2.$$

The (complex) eigenvalues of G are $(\mathrm{sech}\beta \pm i\tanh\beta)^2$, with corresponding eigenvector given by $(\mp i, e^{\beta})^T$. We can now write down the values of α_{2k} and γ_{2k} explicitly: if we define the complex constant C by

$$\begin{pmatrix} \alpha \\ \gamma \end{pmatrix} = C \begin{pmatrix} -i \\ e^{\beta} \end{pmatrix} + C^* \begin{pmatrix} i \\ e^{\beta} \end{pmatrix},$$

then we have

$$\begin{pmatrix} \alpha_{2k} \\ \gamma_{2k} \end{pmatrix} = C(\mathrm{sech}\beta + i\tanh\beta)^{2k} \begin{pmatrix} -i \\ e^{\beta} \end{pmatrix} + C^*(\mathrm{sech}\beta - i\tanh\beta)^{2k} \begin{pmatrix} i \\ e^{\beta} \end{pmatrix}.$$

Depending on our choice of initial 2-wave strength β, there are two possibilities. First, if the eigenvalues are nth roots of unity, so $\mathrm{sech}\beta = \cos\frac{m\pi}{n}$, then $\alpha_{2n} = \alpha$ and $\gamma_{2n} = \gamma$, so that our solutions are periodic in time with period $4n\Delta x$. (There is an extra sign change in the waves.) On the other hand, if the eigenvalues are not roots of unity, there is a continuous transfer of energy between the two wave families and oscillation of wave strengths, but solutions are not time-periodic. Here we view G as a skewed

rotation of the $\alpha - \gamma$ plane, and the orbit of G is dense in an ellipse containing the point (α, γ). In other words, the origin is a nonhyperbolic fixed point for the map G. In particular, the periodic solutions are not stable under perturbations of the 2-wave strength β.

We now consider solutions of the conservation law with flux f_2, for which we took $\tau(v) = 1$ and $\sigma(v) = 2v$. Again we have $v_2 - v_1 = \beta$, and substituting these values into (5.9), we get

$$G(v_1, v_2) = \frac{1}{1-\beta^2} \begin{pmatrix} 1+\beta^2 & 2\beta \\ 2\beta & 1+\beta^2 \end{pmatrix}.$$

The eigenvalues of G are $\frac{1-\beta}{1+\beta}$ and $\frac{1+\beta}{1-\beta}$, with corresponding eigenvectors $(1,-1)^T$ and $(1,1)^T$, respectively. Since these eigenvalues are real, we have one growing mode and one decaying mode. For definiteness, suppose that $0 < \beta < 1$. Then choosing $\alpha = \gamma$ initially is the pure growth mode, and we have

$$\alpha_{2k} = \gamma_{2k} = \left(\frac{1+\beta}{1-\beta}\right)^k \alpha, \tag{5.10}$$

so that the 1- and 3-components of the solution grow exponentially in time. The other extreme is when $\gamma = -\alpha$, in which case we get exponential decay in the first and third families. In general, there will be a combination of growth and decay, but there will be some growth in L^p-norms of the solution to the conservation law, if the projection of the vector $(\alpha, \gamma)^T$ onto the growing mode is nonzero. This indicates that these solutions are more stable under perturbations than those above. In the $\alpha - \gamma$ plane, the origin is a hyperbolic fixed point of the map G corresponding to a saddle.

Since (5.10) gives an exact expression for the waves α_{2k} and γ_{2k}, we can calculate all norms of the solutions directly. Since there is one jump in v of size β per Δx, we take $d = \Delta x$ and we have $\|v\|'_d = \beta = \rho \Delta x$, where, as in section 4, ρ is a scaling parameter which determines the growth rate. In particular, we have for $t = 2k\Delta x$,

$$\begin{aligned} TV\big(u(\cdot,t)\big) &= TV\big(u_0\big) \left(\frac{1+\beta}{1-\beta}\right)^k \\ &= TV\big(u_0\big) \left(\frac{1+\rho\Delta x}{1-\rho\Delta x}\right)^{t/2\Delta x} \\ &\geq TV\big(u_0\big) \exp(\rho t), \end{aligned}$$

since $\frac{1+y}{1-y} \geq e^{2y}$ for $y \leq 1$, which proves the corollary. Indeed, since $\frac{1+y}{1-y} = e^{2y} + O(y^3)$, we can make the growth as close to $e^{\rho t}$ as we like.

For the flux f_3, similar calculations follow. We shall not carry out the tedious details but rather describe the general situation. The matrix $G(v_1, v_2)$ is viewed as a linear map on the $\alpha - \gamma$ plane. It has determinant 1, and the diagonal entries are equal and are given by $\mu = \frac{1}{2}\mathrm{tr}(G)$. As we noted earlier, the eigenvalues η and $1/\eta$ of G, which determine the dynamics of the map, are complex (with unit length) if $\mu^2 < 1$, and real otherwise. For $\tau(v) > 0$, an equivalent description is the following: defining

$$\Sigma(v_1, v_2) = \big(\sigma(v_2)\tau(v_1) - \sigma(v_1)\tau(v_2)\big)^2 - \big(\tau(v_2) - \tau(v_1)\big)^2,$$

we have complex eigenvalues for $\Sigma < 0$ and real eigenvalues for $\Sigma > 0$. As above, complex eigenvalues correspond to oscillations without growth, while real eigenvalues

correspond to growth and decay. Thus, each time Σ changes sign, we get a bifurcation from one type of behavior to another. In particular, for the flux f_3, we have $\sigma(v) = v$ and $\tau(v) = e^v$. Now fixing $v_1 = 0$ and $v_2 = \beta$, we get $\Sigma = \beta^2 - (e^\beta - 1)^2 = -\beta^3 + O(\beta^4)$, which changes sign as β passes through 0. Thus, for $\beta < 0$ we get growth and decay, while for $\beta > 0$ we get oscillations and periodic solutions. □

We remark that in all of the above examples, we obtain less interesting periodic solutions as follows. Referring again to Figure 3.1, we choose the initial 2-waves 12 and 56 to vanish, while choosing 34 and 70 to be β and $-\beta$, respectively, and choosing the 1- and 3-wave strengths as above. Calculations similar to the above then show that the solutions with these data are indeed periodic with period $d = 4\Delta x$, irrespective of the flux which is used.

Since these solutions are exact and global, they do not depend on the size of the initial data, and we can again scale the parameter Δx and the initial wave strengths to change the quantitative behavior of solution to get different values of d and the d-norm. Thus the rate at which the solution grows or decays, or the period of the solution, can be set according to our needs. These simple examples indicate that there is a wide variety of nonlinear phenomena associated with these simple equations. The author and Blake Temple have recently shown that for general systems with one Riemann coordinate, if the sup-norm of the data is small enough, then there can be at most exponential growth in the solution, with growth rate depending on d [10]. The examples presented here show that this result is sharp in the absence of further assumptions on the flux.

Appendix. Construction of the flux. We now describe the systems of degenerate conservation laws which satisfy the conditions used in earlier sections. Recall that the conservation law is given by

$$u_t + f(u)_x = 0,$$

where $u \in \mathbf{R}^3$ and $f : \mathcal{U} \to \mathbf{R}^3$ is a smooth vector valued function in an open neighborhood $\mathcal{U} \subset \mathbf{R}^3$. The hyperbolic wave-speeds are given by the (constant) eigenvalues λ_i of the matrix $A(u) = Df_u$, and the corresponding wave curves are the integrals of the right eigenvectors $r^i(u)$. The local nonlinear interactions of waves of different families are determined by the Lie algebra of eigenvector fields, quantified by the interaction coefficients

$$\Lambda_i \equiv \Lambda_i^{jk} = l_i \cdot [r^j, r^k], \quad \text{where} \quad i \neq j > k \neq i,$$

evaluated at the single point $\tilde{u} \in \mathcal{U}$, which we take to be the origin.

Our task is thus to find a matrix $A(u)$ with constant (distinct) eigenvalues and whose eigenvectors are such that the coefficients Λ_i^{jk} take on specified values at the origin. In addition, the matrix $A(u)$ should be the gradient of some vector function defined in a neighborhood of the origin. When there was a Riemann coordinate, we wrote down the system in closed form; if there is no Riemann coordinate we are not able to do this.

The eigenvalues λ_i of A are the roots of the characteristic polynomial $P_A(\lambda) = \det(A - \lambda I)$ of $A = [A_{ij}]$. The condition that the eigenvalues of $A(u)$ be constant in a neighborhood is thus equivalent to the statement that $P_A(\lambda)$ does not depend on u. In addition, we require that the curl of each row vanishes, so that each row of A is the gradient of a scalar function.

We reformulate the problem as a system of PDEs for the components $A_{ij}(u)$ with independent variables u_k, as follows. We shall consider a Cauchy problem, where

we take the Cauchy surface to be the plane $u_1 = 0$, and describe changes in the components A_{ij} of A in the (time-like) direction of u_1.

Since the characteristic polynomial $P_A(\lambda)$ is to be constant in the neighborhood, differentiating (each coefficient) with respect to u_1 must give zero. This gives three nonlinear equations, namely one for each coefficient of $P_A(\lambda)$. Since we have nine dependent variables, we need six more equations. These are obtained using the requirement that the rows of A be gradients, or curl-free. This is the condition that

$$\partial A_{ij}/\partial u_k = \partial A_{ik}/\partial u_j, \tag{A.1}$$

which is a nontrivial restriction for each $j \neq k$. We wish to describe changes in the time-like direction u_1, so we take as the remaining equations those obtained by letting $k = 1$ and $j = 2$ or 3, while i varies from 1 to 3. Thus our full system of equations is given by

$$\begin{aligned} &\partial P_A(\lambda)/\partial u_1 = 0 \quad \text{and} \\ &\partial A_{ij}/\partial u_1 = \partial A_{i1}/\partial u_j. \end{aligned} \tag{A.2}$$

If we choose analytic initial data, the Cauchy–Kowalewski theorem implies the existence and uniqueness of a local solution to this system. It remains to choose initial data on a noncharacteristic surface in such a way that the solution matrix $A = [A_{ij}]$ satisfies our requirements.

We take the Cauchy surface to be the plane $u_1 = 0$ and choose data A^0 satisfying the above conditions on the plane. That is, A^0 should have constant eigenvalues, while equations (A.1) hold in the plane $u_1 = 0$, when u_2 and u_3 vary. Uniqueness of solutions then implies that since $P_A(\lambda)$ is constant on the plane $u_1 = 0$ and satisfies $P_A(\lambda)_{u_1} = 0$, it is constant everywhere. Similarly, differentiating the linear equations in (A.2) gives

$$\frac{\partial^2 A_{i2}}{\partial u_3 \partial u_1} = \frac{\partial^2 A_{i1}}{\partial u_3 \partial u_2} = \frac{\partial^2 A_{i3}}{\partial u_2 \partial u_1}$$

for each i. Thus, again by uniqueness, if the Cauchy data satisfies (A.1) with $j = 2$ and $k = 3$, then this is satisfied throughout the neighborhood, and $A(u)$ is indeed a gradient.

It remains to choose appropriate Cauchy data $A^0(u_2, u_3)$ and calculate the interaction coefficients. We shall choose the second and third columns of A^0 to satisfy (A.1) and then choose the first column so that we get the correct characteristic polynomial. The interaction coefficients are given in terms of the gradient DA of A by the formula

$$\Lambda_i^{jk} = \frac{\lambda_j - \lambda_k}{(\lambda_j - \lambda_i)(\lambda_k - \lambda_i)} l_i \cdot DA(r^j) \cdot r^k, \tag{A.3}$$

so we calculate DA at the origin [11]. This amounts to finding all partial derivatives of the A_{ij} at the origin. These are found from (A.2) as follows. The quantities $\partial A_{ij}/\partial u_k$ can be immediately calculated for all i and j and for $k = 2$ and 3. The derivatives $\partial A_{ik}/\partial u_1$, for $k = 2$ and 3, are then equated with $\partial A_{i1}/\partial u_k$. Finally, $\partial A_{i1}/\partial u_1$ are obtained implicitly from the equation $P_A(\lambda)_{u_1} = 0$, by substituting in the previously found values for $\partial A_{ik}/\partial u_1$.

We carry out the details, assuming that the eigenvalues are -1, ν, and 1. Given some function $\psi(u_2, u_3)$ to be chosen later, define the matrix

$$A^0(u_2, u_3) = \begin{pmatrix} \alpha & 1 & 0 \\ \beta & 0 & 1 \\ \gamma & \psi_{u_2} & \psi_{u_3} \end{pmatrix},$$

where we must solve for α, β, and γ as functions of u_2 and u_3 so that the correct characteristic polynomial is realized. The coefficients of $P_{A^0}(\lambda)$ are easily seen to be $\alpha + \psi_{u_3}$, $\beta + \psi_{u_2} - \alpha\psi_{u_3}$, and $\gamma - \alpha\psi_{u_2} - \beta\alpha\psi_{u_3}$, respectively. It is now clear what α, β, and γ should be as functions of u_2 and u_3 to ensure a given constant characteristic polynomial on the plane $u_1 = 0$. If we make the further assumption that the first partial derivatives of ψ vanish at the origin, then it is easy to calculate the eigenvectors there. Indeed, we have

$$A^0(0,0) = \begin{pmatrix} \alpha_0 & 1 & 0 \\ \beta_0 & 0 & 1 \\ \gamma_0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} \nu & 1 & 0 \\ 1 & 0 & 1 \\ -\nu & 0 & 0 \end{pmatrix},$$

whose eigenvectors are easily calculated.

To calculate DA at the origin, we differentiate the characteristic polynomial and evaluate at the origin, to get the system of three equations

$$\begin{aligned} A_{11,z} + A_{22,z} + A_{33,z} &= 0, \\ \alpha_0(A_{22,z} + A_{33,z}) - A_{21,z} - \beta_0 A_{12,z} - \gamma_0 A_{13,z} - A_{32,z} &= 0, \quad \text{and} \\ \gamma_0(A_{12,z} + A_{23,z}) + A_{31,z} - \alpha_0 A_{32,z} - \beta_0 A_{33,z} &= 0. \end{aligned} \tag{A.4}$$

We now take the dummy variable z to be u_2 and u_3 to get the derivatives of A_{i1} with respect to these; use symmetry; and finally take z to be u_1 to get the $\partial A_{i1}/\partial u_1$. Denoting $\partial A_{ij}/\partial u_k$ by $A_{ij,k}$ and $\psi_{u_j u_k}$ by ψ_{jk}, using (A.4) with $z = u_2$ and u_3 together with symmetry, we find

$$\begin{aligned} A_{12,1} &= A_{11,2} = -\psi_{23}, \\ A_{13,1} &= A_{11,3} = -\psi_{33}, \\ A_{22,1} &= A_{21,2} = \nu\psi_{23} - \psi_{22}, \\ A_{23,1} &= A_{21,3} = \nu\psi_{33} - \psi_{23}, \\ A_{32,1} &= A_{31,2} = \nu\psi_{22} + \psi_{23}, \quad \text{and} \\ A_{33,1} &= A_{31,3} = \nu\psi_{23} + \psi_{33}. \end{aligned}$$

Finally, we substitute these values back into (A.4) with $z = u_1$ to get

$$\begin{aligned} A_{11,1} &= \psi_{22} - 2\nu\psi_{23} - \psi_{33}, \\ A_{21,1} &= 2\nu^2\psi_{23} - 2\nu\psi_{22}, \quad \text{and} \\ A_{31,1} &= \nu^2\psi_{22} + (1+\nu^2)\psi_{33}. \end{aligned}$$

This determines completely the total derivative $DA(0,0,0)$ of A at the origin, and allows us to calculate the interaction coefficients. Note that (A.4) with $z = u_1$ determines $A_{i1,1}$ at the origin which, together with the linear equations in (A.2), shows that the Cauchy surface $u_1 = 0$ is noncharacteristic.

The eigenvectors at the origin are

$$\begin{aligned} r^1 &= (1 \quad -1-\nu \quad \nu)^T, \\ r^2 &= (1 \quad 0 \quad 1)^T, \qquad\qquad \text{and} \\ r^3 &= (1 \quad 1-\nu \quad -\nu)^T, \end{aligned}$$

respectively, and we take $\psi(u_2, u_3)$ to be quadratic with second derivatives

$$\begin{aligned} \psi_{22} &= \frac{2}{1-\nu^2} \pm \frac{1}{2}\nu^2, \\ \psi_{23} &= \frac{-2\nu}{1-\nu^2} \pm \frac{1}{2}\nu, \quad \text{and} \\ \psi_{33} &= \frac{2}{1-\nu^2} \pm \frac{1}{2}. \end{aligned}$$

Substituting all of the above into (A.3), we calculate the coefficients to be $\{1, \pm 1, 1\}$. Thus we have found a matrix $A(u)$ that satisfies the necessary conditions and that has these interaction coefficients.

REFERENCES

[1] J. Glimm, *Solutions in the large for nonlinear hyperbolic systems of equations*, Comm. Pure Appl. Math., 18 (1965), pp. 697–715.

[2] J. Glimm and P. Lax, *Decay of solutions of systems of nonlinear hyperbolic conservation laws*, Mem. Amer. Math. Soc., 101 (1970).

[3] J. M. Greenberg, *Smooth and time-periodic solutions to the quasilinear wave equation*, Arch. Rational Mech. Anal., 60 (1975), pp. 29–50.

[4] J. M. Greenberg and M. Rascle, *Time-periodic solutions to systems of conservation laws*, Arch. Rat. Mech. Anal., 115 (1991), pp. 395–407.

[5] J. Hunter, *Strongly nonlinear hyperbolic waves*, in Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications, J. Ballmann and R. Jeltsch, eds., Vieweg, 1989, pp. 257–268.

[6] J.-L. Joly, G. Métivier, and J. Rauch, *A nonlinear instability for* 3×3 *systems of conservation laws*, Comm. Math. Physics, 162 (1994), pp. 47–59.

[7] P. Lax, *Hyperbolic systems of conservation laws*, II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[8] 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.

[9] A. Majda and R. Rosales, *Resonantly interacting weakly nonlinear hyperbolic waves* I. *A single space variable*, Stud. Appl. Math., 71 (1984), pp. 149–179.

[10] B. Temple and R. Young, *The large time stability of sound waves*, Comm. Math. Phys., 179 (1996), pp. 417–466.

[11] R. Young, *On Elementary Interactions for Hyperbolic Conservation Laws*, preprint, 1994.

[12] R. Young, *Sup-norm stability for Glimm's scheme*, Comm. Pure Appl. Math., 46 (1993), pp. 903–948.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 559–583

SCATTERING OF ELECTROMAGNETIC WAVES BY ROUGH INTERFACES AND INHOMOGENEOUS LAYERS*

SIMON N. CHANDLER-WILDE† AND BO ZHANG‡

Abstract. We consider a two-dimensional problem of scattering of a time-harmonic electromagnetic plane wave by an infinite inhomogeneous conducting or dielectric layer at the interface between semi-infinite homogeneous dielectric half-spaces. The magnetic permeability is assumed to be a fixed positive constant. The material properties of the media are characterized completely by an index of refraction, which is a bounded measurable function in the layer and takes positive constant values above and below the layer, corresponding to the homogeneous dielectric media. In this paper, we examine only the transverse magnetic (TM) polarization case. A radiation condition appropriate for scattering by infinite rough surfaces is introduced, a generalization of the Rayleigh expansion condition for diffraction gratings. With the help of the radiation condition the problem is reformulated as an equivalent mixed system of boundary and domain integral equations, consisting of second-kind integral equations over the layer and interfaces within the layer. Assumptions on the variation of the index of refraction in the layer are then imposed which prove to be sufficient, together with the radiation condition, to prove uniqueness of solution and nonexistence of guided wave modes. Recent, general results on the solvability of systems of second kind integral equations on unbounded domains establish existence of solution and continuous dependence in a weighted norm of the solution on the given data. The results obtained apply to the case of scattering by a rough interface between two dielectric media and to many other practical configurations.

Key words. scattering, integral equation, inhomogeneous medium, Helmholtz equation

AMS subject classifications. 35J05, 35L05, 45E10, 78A45

PII. S0036141097328932

1. Introduction.

Consider a time harmonic electromagnetic plane wave incident on a layer of some inhomogeneous, isotropic, conducting, or dielectric material in $\mathbf{R}^3$. The media, above and below the layer, consist of some homogeneous dielectric materials. Adopting Cartesian axes $0x_1x_2x_3$, we assume throughout that the material is invariant in the x_3 direction. Thus, in effect, the problem geometry is two-dimensional. Further, we assume that the magnetic permeability is a fixed positive constant. The material properties of the media are then characterized completely by an index of refraction, dependent on the permittivity and conductivity, which is assumed to be a bounded measurable function in the layer and takes positive constant values above and below the layer. The scattering problem is to study the electromagnetic field distributions.

In this paper we formulate first the scattering problem as a boundary value problem for the reduced wave equation (Helmholtz equation), using a radiation condition recently introduced for problems of scattering by infinite one-dimensional rough surfaces and interfaces [4, 5, 7, 8, 9], which is a generalization of the usual radiation condition used in the study of plane wave diffraction by one-dimensional periodic gratings (see, e.g., [23, 1, 2, 4, 19, 20, 25]). Next, in section 3, we derive a novel

*Received by the editors October 20, 1997; accepted for publication (in revised form) July 27, 1998; published electronically March 29, 1999. This work was supported by the UK Engineering and Physical Sciences Research Council under grant GR/K24406.

http://www.siam.org/journals/sima/30-3/32893.html

†Department of Mathematics and Statistics, Brunel University, Uxbridge UB8 3PH, UK (Simon. Chandler-Wilde@ brunel.ac.uk).

‡School of Mathematical and Information Sciences, Coventry University, Coventry CV1 5FB UK (b.zhang@coventry.ac.uk).

integral equation formulation of the problem, as a system of coupled second-kind domain and boundary integral equations, over the layer and over interfaces within the layer, and establish that this formulation is equivalent to the formulation as a boundary value problem. The radiation condition imposed does not rule out guided waves, localized in the inhomogeneous layer, which are thus solutions of the homogeneous boundary value problem and the homogeneous integral equation formulation. From section 4 onward we make restrictions on the variation of the index of refraction in the layer. Under these restrictions we establish, in section 4, an a priori inequality satisfied by any solution. Using this inequality, a key lemma from [8], and extensions of arguments in [7, 31], uniqueness results and hence conditions for the nonexistence of guided wave modes are established in section 5. In section 6, existence of solution is established by employing a novel form of Fredholm alternative based on general results on the solvability of systems of integral equations on unbounded domains in [10].

The assumptions we impose on the index of refraction from section 4 onward are satisfied in many practical cases. In particular, the results obtained apply to the case of scattering by a rough interface between two dielectric media and apply to scattering by a homogeneous layer having rough interfaces, with the media above and below, provided that the wavenumbers in the layer (k^*), and in the media above (k_+) and below (k_-), satisfy either that $\Im k^* > 0$ or that $\max(k_-, k_+) > k^*$. For a precise statement of the cases covered see section 2. Our conclusion that no guided waves exist if $\max(k_-, k_+) > k^*$ or $\Im k^* > 0$ is in agreement with explicit analytical calculations of guided wave modes for the case of plane interfaces between the layer and the media [24].

Integral equation methods have been used widely in the theoretical and numerical study of wave scattering by finite obstacles or local inhomogeneities (see, e.g., [12, 13] and the references quoted there). More recently they have been employed to study scattering by periodic structures [11, 17, 19, 20, 22] and by a nonstratified local inhomogeneity in a stratified medium [30]. Integral equation formulations have also been used extensively in computations of wave scattering by infinite one- and two-dimensional rough surfaces and interfaces (see, e.g., [26, 14, 21, 29] and the references quoted there), but little attention appears to have been paid in the literature to their mathematical justification (a recent exception is [15]).

This present paper is intended, in part, as a contribution to the mathematical analysis of rough surface scattering problems and of the well-posedness of their formulation as integral equations. It is related, in terms of results and methods of argument, to recent studies of scattering of a wave incident from a homogeneous half-space onto an inhomogeneous impedance plane [5]; of electromagnetic waves by a one-dimensional perfectly conducting rough surface [6, 8]; of electromagnetic waves by an inhomogeneous conducting or dielectric layer on a perfectly conducting plate [7]; and of acoustic waves by an inhomogeneous layer on a rigid plate [31]. In particular, the present study is closest to this last paper [31] in that the existence proof depends on the same novel results on the solvability of systems of weakly singular second-kind integral equations on unbounded domains. However, the whole space problem considered here requires a substantially more elaborate uniqueness proof and integral equation formulation, related to the presence of transmitted as well as reflected waves, in contrast to the half-plane problems considered in [5, 6, 8, 7, 31]. Moreover, in these latter papers integral equation formulations using half-plane Green's functions appear natural: as proposed here, the formulation of the whole-space problem as a system of integral

equations in overlapping half-planes, using half-plane Green's functions, is surprising but proves powerful in establishing uniqueness and existence results.

This paper can also be viewed as a generalization of the results of Bonnet-Bendhia and Starling [3] and Strycharz [25], who study plane wave scattering by an inhomogeneous *periodic* layer. In fact, our uniqueness results derive in part from those of Bonnet-Bendhia and Starling [3] and Strycharz [25], and include some of their results for a periodic layer as special cases; however, note that our results are obtained without an a priori assumption of quasi periodicity of the scattered field. We note also that our existence arguments, based on integral equation methods, differ from the variational methods used in [3] and [25] which appear restricted to the periodic case. We further point out that while integral equation-based existence proofs are common (and more straightforward) in the periodic, diffraction grating case (e.g., [19, 11]), they usually fail for a discrete set of combinations of grating period and angle of incidence at which the integral equation formulation is undefined. Our results show that, at least in the two-dimensional case, this problem can be avoided by use of a half-plane rather than a whole-space Green's function in the integral equation formulation. A finite element method for the case of plane wave scattering by an inhomogeneous periodic layer is analyzed in [1].

We remark that aside from the theoretical use to which the formulation is put in this paper, we anticipate that the novel integral equation formulation we derive in section 3 may also be of value for numerical computation. We point out that the integral operators in our formulation are exclusively of convolution type or are products of convolution and multiplication operators so that, after discretization, the matrix vector multiplications required in an iterative solution scheme can be performed efficiently using the FFT (see [27, pp. 109–111] and [28]). We further point out that, even for the simple case of a single interface between two dielectric media, for which a boundary integral equation formulation on the interface is usual (avoiding domain integrals), a recently successful numerical algorithm involves imbedding the one-dimensional boundary curve in a two-dimensional grid so that FFT techniques can be applied [29].

We conclude this section by introducing some notations used throughout. For $h \in \mathbf{R}$, define $\Gamma_h = \{x = (x_1, x_2) \in \mathbf{R}^2 | x_2 = h\}$ and $U_h = \{x \in \mathbf{R}^2 | x_2 > h\}$. Set $E_h^H = U_h \backslash \overline{U}_H$ for $H > h$, and write U, Γ, and E_H for U_0, Γ_0, and E_0^H, respectively. Define $D_A = \{x \in \mathbf{R}^2 | \, |x_1| < A\}$, $A > 0$, and $\Gamma_h(A) = \Gamma_h \cap D_A$, $E_h^H(A) = E_h^H \cap D_A$. For $G \subset \mathbf{R}^2$ let $BC(G)$ denote the space of bounded continuous functions defined on G. For $v \in C^1(\mathbf{R}^2)$ denote by $\partial_j v$, $j = 1, 2$, the derivative $\partial v(x)/\partial x_j$. Finally, for $A > 0$, $x \in \mathbf{R}^2$, let $B_A(x) = \{y \in \mathbf{R}^2 | \, |y - x| < A\}$.

2. The scattering problem and radiation conditions. Let us assume that $\mathbf{R}^3$ is filled with an inhomogeneous, isotropic, conducting, or dielectric medium of electric permittivity $\epsilon > 0$, magnetic permeability $\mu > 0$, and electric conductivity $\sigma \geq 0$. Suppose that the medium is nonmagnetic, i.e., the magnetic permeability μ is a fixed constant in $\mathbf{R}^3$, and suppose that the fields are source free. Then the electromagnetic wave propagation is governed by the time-harmonic Maxwell equations (time dependence $\exp(-i\omega t)$ with frequency $\omega > 0$)

$$\nabla \times E - i\omega\mu H = 0, \tag{2.1}$$

$$\nabla \times H + (i\omega\epsilon - \sigma)E = 0, \tag{2.2}$$

where E and H are the electric field and magnetic field, respectively. In this paper, it is assumed that the medium is invariant in the x_3 direction, i.e., $\epsilon = \epsilon(x)$ and $\sigma = \sigma(x)$

with $x = (x_1, x_2) \in \mathbf{R}^2$. Also, we restrict ourselves to the transverse magnetic (TM) polarization case; that is, the electric field E is assumed to point along the x_3 axis. Let $E = (0, 0, u)$, where $u = u(x)$ is a scalar function. Then it follows from the Maxwell equations (2.1)–(2.2) that u satisfies the reduced wave equation

$$\Delta u + k^2 u = 0 \quad \text{in} \quad \mathbf{R}^2, \tag{2.3}$$

where Δ is the Laplacian in $\mathbf{R}^2$ and $k^2 = \omega^2 \mu\epsilon[1 + i\sigma/(\omega\epsilon)]$ so that $\Im(k^2) \geq 0$.

Additionally we make the following assumptions on k throughout:

(A1) $k \in L_\infty(\mathbf{R}^2)$.

(A2) There are positive constants B, k_+, and k_- such that $k(x) = k_+$ for $x \in U_B$, $= k_-$ for $x \in \mathbf{R}^2 \backslash \overline{U}$.

These two assumptions are sufficient (together with the radiation conditions we introduce below) to derive, in section 3, an equivalent integral equation formulation of the problem. In sections 4 and 5 we address the question of uniqueness of solution which is related to the question of existence or otherwise of guided wave solutions of the homogeneous problem.

We remark that the radiation conditions we will impose will ensure that the scattered field does not contain a downward propagating component and that the transmitted wave does not contain an upward propagating component but (in common with the usual radiation condition for plane wave incidence on periodic gratings) will not rule out solutions of the homogeneous problem which are guided waves localized in the inhomogeneous layer. (See Theorem A.1 in the appendix, where a precise definition of a guided wave in this context is given.) Thus, to prove any uniqueness result, we will have to impose additional conditions (on k) which rule out guided waves. In other words (and more positively), any uniqueness proof will simultaneously establish conditions for the nonexistence of guided waves.

The additional requirements for our uniqueness proof (sections 4 and 5) and for proving the existence of solution (section 6) are that assumption (A3) is satisfied or that both assumptions (A4) and (A5) below are satisfied.

(A3) There exist constants $\lambda_1, \lambda_2, \eta$, and ρ, with $\lambda_1 > 0$, $1 > \lambda_2 > 0$ and $0 \leq \eta < \rho \leq B$ such that $\Im(k^2(x)) \geq \lambda_1$ for almost all $x \in E_\eta^\rho$, $\Im(k^2(x)) \geq \lambda_2 |k^2(x) - k_+^2|$ for almost all $x \in E_\eta^B$, and $\Im(k^2(x)) \geq \lambda_2 |k^2(x) - k_-^2|$ for almost all $x \in E_0^\rho$.

(A4) There exists $\beta \in \mathbf{R}$ such that $\Re(k^2(x))$ is monotonic nondecreasing in U_β and monotonic nonincreasing in $\mathbf{R}^2 \backslash U_\beta$ as x_2 increases: precisely, for all $h > 0$, where $e_2 = (0, 1)$, $\Re[k^2(x + e_2 h)] \geq \Re[k^2(x)]$ for almost all $x \in U_\beta$ and $\Re[k^2(x - e_2 h)] \geq \Re[k^2(x)]$ for almost all $x \in \mathbf{R}^2 \backslash \overline{U}_\beta$.

Let $\tilde{k}(x) = k_+$ for $x_2 \geq \beta$, $= k_-$ for $x_2 < \beta$. Then Assumption (A4) implies that $\Re[k^2(x)] \leq \tilde{k}^2(x)$ for almost all $x \in \mathbf{R}^2$.

(A5) There are constants λ_3, η, and ρ, with $\lambda_3 > 0$ and $0 \leq \eta < \rho \leq B$, such that $\Re[k^2(x)] \leq \tilde{k}^2(x) - \lambda_3$ for almost all $x \in E_\eta^\rho$.

We can write (A4) succinctly as

$$(x_2 - \beta)\partial_2(\Re(k^2)) \geq 0$$

in a distributional sense in $\mathbf{R}^2$ (cf. Bonnet-Bendhia and Starling [3, equation (3.34)] and [7]). If (A4) and (A5) hold with $\beta < \rho$, then it must be the case that $\Re[k^2(x)] \leq k_+^2 - \lambda_3$ for almost all $x \in E_\beta^\rho$, while if $\beta > \eta$, then $\Re[k^2(x)] \leq k_-^2 - \lambda_3$ for almost all $x \in E_\eta^\beta$.

To clarify the above assumptions we list some important practical cases in which they are satisfied:

(i) Suppose that $k_+ \neq k_-$ and, for some $f \in L_\infty(\mathbf{R})$, that $k(x) = k_+$, $x_2 > f(x_1)$, $= k_-$, $x_2 < f(x_1)$, and assume without loss of generality that $\epsilon \leq f(x_1) \leq B - \epsilon$ for some $\epsilon > 0$, $B > \epsilon$. Then (A1), (A2), (A4), (A5) are satisfied, with $0 = \beta = \eta$, $\rho = \epsilon$, if $k_- < k_+$, $B - \epsilon = \eta$, $B = \beta = \rho$, if $k_- > k_+$.

(ii) Suppose that $k^* > 0$ and, for some $f_+, f_- \in L_\infty(\mathbf{R})$, with $f_- \leq f_+$, that $k(x) = k_+$, $x_2 > f_+(x_1)$, $= k_-$, $x_2 < f_-(x_1)$, $= k^*$, $f_-(x_1) < x_2 < f_+(x_1)$. Suppose further without loss of generality that, for some $B > \epsilon > 0$, $\epsilon \leq f_-(x_1) \leq f_+(x_1) \leq B - \epsilon$, $x_1 \in \mathbf{R}$. Then assumptions (A1) and (A2) are satisfied and so are assumptions (A4) and (A5) in the following cases: (a) $k_- < k^* < k_+$ (set $\beta = \eta = 0$, $\rho = \epsilon$); (b) $k_- > k^* > k_+$ (set $\eta = B - \epsilon$, $\beta = \rho = B$); (c) $k^* < k_+$, $k^* < k_-$ provided, for some $0 < \eta < \rho < B$, $k(x) = k^*$, $x \in E_\eta^\rho$, and $\eta \leq \beta \leq \rho$.

(iii) Suppose that $\Im k^* > 0$ and that, for some $0 < \eta < \rho < B$, and disjoint open sets S_+, S, and S_-, with $\overline{S_+} \cup \overline{S} \cup \overline{S_+} = \mathbf{R}^2$ and $E_\eta^\rho \subset S$, $U_B \subset S_+$, and $\mathbf{R}^2 \setminus \overline{U} \subset S_-$,

$$k(x) = \begin{cases} k_+, & x \in S_+, \\ k^*, & x \in S, \\ k_-, & x \in S_-. \end{cases}$$

Then (A1)–(A3) are satisfied.

We mention one simple example not covered by the above assumptions. In the case $k \equiv k_+$ assumptions (A1), (A2), and (A4) are satisfied (with $k_- = k_+$), but assumption (A5) is not satisfied. Thus, the uniqueness results established in section 4 do not apply, and indeed, our problem as formulated will not have a unique solution in this case as is shown by the simple example $u(x) = \exp(\pm i k_+ x_1)$, which satisfies (2.3) with $k \equiv k_+$ and, by Remark 2.3 below, the radiation conditions (2.4) and (2.5).

Let $u^{\mathrm{i}}(x) = \exp(ik_+ x \cdot \alpha)$ be the time-harmonic incoming plane wave incident from U_B on the finite inhomogeneous layer E_B, where $x \in \mathbf{R}^2$, $\alpha = (\cos\theta, -\sin\theta) \in \mathbf{R}^2$, and $\theta \in (0, \pi)$ is the incident angle. We are interested in finding the total field u satisfying the reduced wave equation (2.3).

In order to determine the physical solution u, a radiation condition as x_2 tends to infinity has to be imposed on the scattered field $u^{\mathrm{s}} = u - u^{\mathrm{i}}$ in U_B; that is, the scattered field u^{s} should behave as an outgoing wave as $x_2 \to +\infty$. Similarly, the transmitted field u in $\mathbf{R}^2 \setminus U$ should behave as an outgoing wave as $x_2 \to -\infty$. The standard Sommerfeld radiation condition is not appropriate in this context as we cannot expect that the scattered and the transmitted fields will decay at infinity. We will use a radiation condition proposed in [5] and utilized recently in [7, 8] and [31], which we will usefully relate to the Sommerfeld radiation condition. To this end we introduce the following definitions.

DEFINITION 2.1. *Given a domain $G \subset \mathbf{R}^2$ and $k_* > 0$, call $v \in C^2(G) \cap L_\infty(G)$ a* radiating solution of the Helmholtz equation for wavenumber k_* in G *if $\Delta v + k_*^2 v = 0$ in G and*

$$v(x) = O(r^{-1/2}),$$
$$\frac{\partial v(x)}{\partial r} - ik_* v(x) = o(r^{-1/2}),$$

as $r = |x| \to \infty$, uniformly in $x/|x|$.

Let $\Phi(x, y; k_\pm)$ denote the free-space Green's function for $\Delta + k_\pm^2$; that is,

$$\Phi(x, y; k_\pm) = \frac{i}{4} H_0^{(1)}(k_\pm |x - y|), \qquad x, y \in \mathbf{R}^2, \qquad x \neq y,$$

with $H_0^{(1)}$ being the Hankel function of the first kind of order zero.

DEFINITION 2.2. *Given a domain $G \subset \mathbf{R}^2$, say that $v_+ : G \to \mathbf{C}$ satisfies the* upward propagating radiation condition (UPRC) for wavenumber k_+ *in G if, for some $H \in \mathbf{R}$ and $\phi_+ \in L_\infty(\Gamma_H)$, it holds that $U_H \subset G$ and*

$$v_+(x) = 2 \int_{\Gamma_H} \frac{\partial \Phi(x,y;k_+)}{\partial y_2} \phi_+(y) ds(y), \qquad x \in U_H; \tag{2.4}$$

and say that $v_- : G \to \mathbf{C}$ satisfies the downward propagating radiation condition (DPRC) for wavenumber k_- *in G if, for some $h \in \mathbf{R}$ and $\phi_- \in L_\infty(\Gamma_h)$, it holds that $\mathbf{R}^2 \backslash \overline{U}_h \subset G$ and*

$$v_-(x) = -2 \int_{\Gamma_h} \frac{\partial \Phi(x,y;k_-)}{\partial y_2} \phi_-(y) ds(y), \qquad x \in \mathbf{R}^2 \backslash \overline{U}_h. \tag{2.5}$$

Note that the existence of the integrals in (2.4) and (2.5) for arbitrary $\phi_+ \in L_\infty(\Gamma_H)$ and $\phi_- \in L_\infty(\Gamma_h)$ is ensured by the bound which follows from the asymptotic behavior of the Hankel function for small and large argument,

$$\left| \frac{\partial \Phi(x,y;k_\pm)}{\partial y_2} \right| \leq C|x_2 - y_2|(|x-y|^{-2} + |x-y|^{-3/2}), \quad x, y \in \mathbf{R}^2, \quad x \neq y, \tag{2.6}$$

which holds for some constant $C > 0$ dependent only on $k_\pm$.

The next lemma states properties of the upward propagating radiation condition needed later and, in particular, shows that, for $h \in \mathbf{R}$, a radiating solution for wavenumber k_* in U_h ($\mathbf{R}^2 \backslash \overline{U}_h$) satisfies the UPRC (DPRC) for wavenumber k_*. We first remark that the DPRC can be expressed, through reflection, in terms of the UPRC.

Remark 2.1. For $x = (x_1, x_2) \in \mathbf{R}^2$ let $x' = (x_1, -x_2)$, and for $G \subset \mathbf{R}^2$ let $G' = \{x' | x \in G\}$. Then $v_- : G \to \mathbf{C}$ satisfies the DPRC for wavenumber k_* in G if and only if $v_+ : G' \to \mathbf{C}$, given by $v_+(x) = v_-(x')$, $x \in G'$, satisfies the UPRC for wavenumber k_* in G'.

LEMMA 2.1 (see [7, Theorem 2.1]). *Given $H \in \mathbf{R}$ and $v : U_H \to \mathbf{C}$, the following statements are equivalent:*

(i) *$v \in C^2(U_H)$, $v \in L_\infty(U_H \backslash U_a)$ for all $a > H$, $\Delta v + k_+^2 v = 0$ in U_H, and v satisfies the UPRC for wavenumber k_+;*

(ii) *$v \in L_\infty(U_H \backslash U_a)$ for some $a > H$ and v satisfies* (2.4) *for each $h > H$ with $\phi = v|_{\Gamma_h}$;*

(iii) *$v \in C^2(U_H)$, $v \in L_\infty(U_H \backslash U_a)$ for all $a > H$, $\Delta v + k_+^2 v = 0$ in U_H, and for every $h > H$ and radiating solution in U_H, w, such that the restrictions of w and $\partial_2 w$ to Γ_h are in $L_1(\mathbf{R})$, it holds that*

$$\int_{\Gamma_h} \left(v \frac{\partial w}{\partial n} - w \frac{\partial v}{\partial n} \right) ds = 0. \tag{2.7}$$

From Lemma 2.1 and Remark 2.1 we can deduce corresponding characteristics of downward propagating solutions of the Helmholtz equation.

For convenience, we now state a local regularity estimate used throughout the paper.

LEMMA 2.2 (see [18, Theorem 3.9, Lemma 4.1]). *If for some $A > 0$ and $x \in \mathbf{R}^2$ it holds that $v \in L_\infty(B_A(x))$ and $\Delta v = f \in L_\infty(B_A(x))$ (in a distributional sense),*

then $v \in C^1(B_A(x))$ *and*

$$|\nabla v(y)| \le CA^{-1}(||v||_\infty + A^2||f||_\infty), \qquad y \in B_{A/2}(x),$$

where C *is an absolute constant.*

Remark 2.2. A consequence of Lemma 2.2 is that if $\Delta v + k^2 v = 0$ in some region G and $v \in L_\infty(G)$, $k \in L_\infty(G)$, then $v \in C^1(G)$ and ∇v is bounded in every compact subset of G. Further, if the sequence $(v_n) \subset L_\infty(G)$ is uniformly bounded, $\Delta v_n + k_*^2 v_n = 0$ in G for some $k_* \in \mathbf{C}$ and each n, and $v_n(x) \to v(x)$ uniformly on compact subsets of G, then $v \in C^2(G)$ and $\Delta v + k_*^2 v = 0$ in G.

Our problem of scattering of a time-harmonic plane wave by an inhomogeneous layer can now be formulated as the following boundary value problem.

Problem (P). Find $u \in C(\mathbf{R}^2)$ such that (i) u satisfies the reduced wave equation (2.3) in a distributional sense; (ii) u^s and u satisfy the UPRC and DPRC (2.4) and (2.5), respectively; and (iii) u is bounded in E^A_{-A} for every $A > 0$.

Remark 2.3. From (iii) and Lemma 2.2, it follows that $u \in C^1(\mathbf{R}^2) \cap C^2(U_B) \cap C^2(\mathbf{R}^2 \backslash \overline{U})$ and

$$\sup_{x \in E^A_{-A}} [|\nabla u(x)| + |u(x)|] < \infty \tag{2.8}$$

for every $A > 0$. Further, by (2.3) and standard local regularity results [18], we have that $u \in H^2_{\mathrm{loc}}(\mathbf{R}^2)$.

Remark 2.4. The radiation conditions (2.4) and (2.5) are generalizations of the standard radiation conditions for one-dimensional periodic gratings. Precisely, it was proven in [4] that if u^s has the usual representation as a Rayleigh expansion [2, 19, 20] in some U_τ, then it also satisfies (2.4) for all $h > \tau$ and thus satisfies the UPRC. As a consequence, any upward or horizontally propagating plane wave satisfies the UPRC and, by Remark 2.1, any downward or horizontally propagating plane wave satisfies the DPRC.

In what follows we are concerned with deriving an equivalent integral equation formulation of Problem (P) and with establishing unique solvability for Problem (P), employing integral equation methods.

3. An integral equation formulation. For $h \in \mathbf{R}$ let $y'_h = (y_1, 2h - y_2)$ be the image of y in Γ_h and define

$$G^\pm_h(x, y) = \Phi(x, y; k_\pm) - \Phi(x, y'_h; k_\pm), \qquad x, y \in \mathbf{R}^2, \qquad x \ne y.$$

Then $G^\pm_h$ is the Dirichlet Green's function for $\Delta + k^2_\pm$ in the half-planes U_h and $\mathbf{R}^2 \backslash \overline{U}_h$. It follows from [6, Lemma 3.1] that, for some constant C depending only on $k_\pm$ and h,

$$|G^\pm_h(x, y)|, \qquad |\nabla_x G^\pm_h(x, y)|, \qquad |\nabla_y G^\pm_h(x, y)| \le C \frac{(1 + x_2)(1 + y_2)}{|x - y|^{3/2}} \tag{3.1}$$

if $x, y \in U_h$ or $x, y \in \mathbf{R}^2 \backslash \overline{U}_h$ with $|x - y| \ge 1$. On the other hand, from asymptotic properties of the Hankel function it follows that

$$|G^\pm_h(x, y)| \le C\,(1 + |\log|x - y||)\,, \quad |\nabla_x G^\pm_h(x, y)|, \quad |\nabla_y G^\pm_h(x, y)| \le C|x - y|^{-1} \tag{3.2}$$

if $x, y \in U_h$ or $x, y \in \mathbf{R}^2 \backslash \overline{U}_h$ with $|x - y| \le 1$. It follows from (3.1) and (3.2) that, for $0 \le h \le B$,

$$|G^\pm_h(x, y)|, \qquad |\nabla_x G^\pm_h(x, y)|, \qquad |\nabla_y G^\pm_h(x, y)| \le C_b(1 + |x_1 - y_1|)^{-3/2} \tag{3.3}$$

if $y \in \overline{E}_h^B$, $x \in \Gamma_b$, with $b > B$, or if $y \in \overline{E}_0^h$, $x \in \Gamma_b$, with $b < 0$, where C_b depends only on b, B, h, and $k_\pm$.

Let u^{r} denote the upward propagating plane wave $u^{\mathrm{r}}(x) = -\exp(ik_+ x_c' \cdot \alpha)$, $x \in \mathbf{R}^2$.

THEOREM 3.1. *Let u be a solution of Problem* (P), *and let $0 \le c < d \le B$. Then we have*

$$u(x) = u^{\mathrm{i}}(x) + u^{\mathrm{r}}(x) + \int_{E_c^B} u(y)[k^2(y) - k_+^2]G_c^+(x,y)dy$$

$$+ \int_{\Gamma_c} u(y)\frac{\partial G_c^+(x,y)}{\partial y_2}ds(y), \qquad x \in U_c, \tag{3.4}$$

$$u(x) = \int_{E_0^d} u(y)[k^2(y) - k_-^2]G_d^-(x,y)dy$$

$$- \int_{\Gamma_d} u(y)\frac{\partial G_d^-(x,y)}{\partial y_2}ds(y), \qquad x \in \mathbf{R}^2 \backslash \overline{U}_d. \tag{3.5}$$

Remark 3.1. In view of (3.1) and (3.2), (A1), and the fact that $u \in BC(\overline{E_B})$, the integrals in (3.4) and (3.5) are well defined.

Proof. Take $x \in U_c$, choose $b > \max(x_2, B)$, $A > |x_1|$, and $\epsilon > 0$ sufficiently small and apply Green's second theorem to $G_c^+(x,\cdot)$ and u in the bounded region $E_c^b(A) \setminus \overline{B_\epsilon(x)}$, and then let $\epsilon \to 0$ to obtain that

$$u(x) = \int_{E_c^B(A)} u(y)[k^2(y) - k_+^2]G_c^+(x,y)dy$$

$$+ \int_{\partial(E_c^b(A))} \left[G_c^+(x,y)\frac{\partial u}{\partial n}(y) - u(y)\frac{\partial G_c^+(x,y)}{\partial n(y)} \right] ds(y). \tag{3.6}$$

Letting $A \to \infty$ in (3.6), in view of (3.1), we find that

$$u(x) = \int_{E_c^B} u(y)[k^2(y) - k_+^2]G_c^+(x,y)dy + \int_{\Gamma_c} u(y)\frac{\partial G_c^+(x,y)}{\partial y_2}ds(y) + I_b, \tag{3.7}$$

where

$$I_b = \int_{\Gamma_b} \left[G_c^+(x,y)\frac{\partial u}{\partial n}(y) - u(y)\frac{\partial G_c^+(x,y)}{\partial n(y)} \right] ds(y). \tag{3.8}$$

Now $v = u^{\mathrm{i}} + u^{\mathrm{r}}$ satisfies the Helmholtz equation $\Delta v + k_+^2 v = 0$ in U_c and the Dirichlet condition $v = 0$ on Γ_c, so that by the same argument used to derive (3.7), we can show that $v(x) = \tilde{I}_b$, where $\tilde{I}_b$ is given by (3.8) but with u replaced by v. It follows that

$$I_b = u^{\mathrm{i}}(x) + u^{\mathrm{r}}(x) + \int_{\Gamma_b} \left[G_c^+(x,y)\frac{\partial w}{\partial n}(y) - w(y)\frac{\partial G_c^+(x,y)}{\partial n(y)} \right] ds(y), \tag{3.9}$$

where $w = u^{\mathrm{s}} - u^{\mathrm{r}}$. Further, by Remark 2.4, u^{r} and thus w satisfies the UPRC. Also, $G_c^+(x,\cdot)$ is a radiating solution in U_τ for $\tau > \max(x_2, B)$ so that, in view of (3.1) and

the equivalence of (i) and (iii) in Lemma 2.1, the integral in (3.9) vanishes. Thus (3.4) follows.

Take $x \in \mathbf{R}^2\backslash\overline{U}_d$, and choose $a < \min(x_2, 0)$, $A > |x_1|$, and $\epsilon > 0$ sufficiently small. Then (3.5) can be derived similarly by applying Green's second theorem to $G_d^-(x,\cdot)$ and u in the bounded region $E_a^d(A) \setminus \overline{B_\epsilon(x)}$, letting $\epsilon \to 0$, $A \to \infty$, and finally utilizing the equivalence of (i) and (iii) in Lemma 2.1 and noting Remark 2.1. □

The next two lemmas state properties of volume and surface potentials of the type appearing in (3.4) and (3.5). Lemma 3.1(i) was proved as Lemma 3.1 in [7] while Lemma 3.2(i) was proved as Theorem 3.2 in [5]. In both lemmas the assertion (ii) is a consequence of (i) on noting Remark 2.1.

LEMMA 3.1. (i) *Define the* volume potential v_+ *with density* $\phi_+ \in L_\infty(E_c^B)$ *by*

$$v_+(x) = \int_{E_c^B} G_c^+(x,y)\phi_+(y)dy, \qquad x \in \overline{U}_c,$$

and extend the definition of ϕ_+ *to* U_c *by setting* $\phi_+(x) = 0$, $x \in U_B$. *Then* $v_+ \in C^1(\overline{U}_c) \cap L_\infty(E_c^b)$ *for* $b > c$, $v_+ = 0$ *on* Γ_c, $\Delta v_+ + k_+^2 v_+ = -\phi_+$ *in* U_c, *and* v_+ *satisfies the UPRC.*

(ii) *Define the* volume potential v_- *with density* $\phi_- \in L_\infty(E_0^d)$ *by*

$$v_-(x) = \int_{E_0^d} G_d^-(x,y)\phi_-(y)dy, \qquad x \in \mathbf{R}^2\backslash U_d,$$

and extend the definition of ϕ_- *to* $\mathbf{R}^2\backslash\overline{U}_d$ *by setting* $\phi_-(x) = 0$, $x \in \mathbf{R}^2\backslash U$. *Then* $v_- \in C^1(\mathbf{R}^2\backslash U_d) \cap L_\infty(E_a^d)$ *for* $a < d$, $v_- = 0$ *on* Γ_d, $\Delta v_- + k_-^2 v_- = -\phi_-$ *in* $\mathbf{R}^2\backslash\overline{U}_d$, *and* v_- *satisfies the DPRC.*

LEMMA 3.2. (i) *Define the* double layer potential D_+ *with density* $\psi_+ \in BC(\Gamma_c)$ *by*

$$D_+(x) = \int_{\Gamma_c} \frac{\partial G_c^+(x,y)}{\partial y_2}\psi_+(y)ds(y), \qquad x \in U_c.$$

Then $D_+ \in C(\overline{U}_c) \cap C^2(U_c) \cap L_\infty(E_c^b)$ *for* $b > c$, $D_+ = \psi_+$ *on* Γ_c, $\Delta D_+ + k_+^2 D_+ = 0$ *in* U_c, *and* D_+ *satisfies the UPRC.*

(ii) *Define the* double layer potential D_- *with density* $\psi_- \in BC(\Gamma_d)$ *by*

$$D_-(x) = \int_{\Gamma_d} \frac{\partial G_d^-(x,y)}{\partial y_2}\psi_-(y)ds(y), \qquad x \in \mathbf{R}^2\backslash\overline{U}_d.$$

Then $D_- \in C(\mathbf{R}^2\backslash U_d) \cap C^2(\mathbf{R}^2\backslash\overline{U}_d) \cap L_\infty(E_a^d)$ *for* $a < d$, $D_- = -\psi_-$ *on* Γ_d, $\Delta D_- + k_-^2 D_- = 0$ *in* $\mathbf{R}^2\backslash\overline{U}_d$, *and* D_- *satisfies the DPRC.*

As in Theorem 3.1, choose c and d so that $0 \le c < d \le B$ and let $\lambda = (k_+ c + k_- d)/(k_- + k_+)$ so that $k_-(d-\lambda) = k_+(\lambda - c)$. Suppose that u satisfies Problem (P) and let $\psi_1 = u|_{\overline{E}_\lambda^B}$, $\psi_2 = u|_{\overline{E}_0^\lambda}$, and $k^\pm = k^2 - k_\pm^2$. Then, by Theorem 3.1, $\psi_1 \in BC(\overline{E}_\lambda^B)$ and $\psi_2 \in BC(\overline{E}_0^\lambda)$ satisfy the pair of second-kind integral equations

$$\psi_1(x) = u^i(x) + u^r(x) + \int_{E_\lambda^B} \psi_1(y)k^+(y)G_c^+(x,y)dy + \int_{E_c^\lambda} \psi_2(y)k^+(y)G_c^+(x,y)dy$$

$$+\int_{\Gamma_c}\psi_2(y)\frac{\partial G_c^+(x,y)}{\partial y_2}ds(y),\qquad x\in\overline{E}_\lambda^B, \tag{3.10}$$

$$\psi_2(x)=\int_{E_0^\lambda}\psi_2(y)k^-(y)G_d^-(x,y)dy+\int_{E_\lambda^d}\psi_1(y)k^-(y)G_d^-(x,y)dy$$

$$-\int_{\Gamma_d}\psi_1(y)\frac{\partial G_d^-(x,y)}{\partial y_2}ds(y),\qquad x\in\overline{E}_0^\lambda. \tag{3.11}$$

Conversely, suppose now that $\psi_1\in BC(\overline{E}_\lambda^B)$ and $\psi_2\in BC(\overline{E}_0^\lambda)$ satisfy the integral equations (3.10) and (3.11) and define u as follows:

$$u(x)=u^{\mathrm{i}}(x)+u^{\mathrm{r}}(x)+\int_{E_\lambda^B}\psi_1(y)k^+(y)G_c^+(x,y)dy+\int_{E_c^\lambda}\psi_2(y)k^+(y)G_c^+(x,y)dy$$

$$+\int_{\Gamma_c}\psi_2(y)\frac{\partial G_c^+(x,y)}{\partial y_2}ds(y),\qquad x\in U_c, \tag{3.12}$$

$$u(x)=\lim_{y\to x,\ y\in U_c}u(y),\qquad x\in\Gamma_c, \tag{3.13}$$

$$u(x)=\int_{E_0^\lambda}\psi_2(y)k^-(y)G_d^-(x,y)dy+\int_{E_\lambda^d}\psi_1(y)k^-(y)G_d^-(x,y)dy$$

$$-\int_{\Gamma_d}\psi_1(y)\frac{\partial G_d^-(x,y)}{\partial y_2}ds(y),\qquad x\in\mathbf{R}^2\backslash\overline{U}_c. \tag{3.14}$$

Then it follows, provided $d-c$ is small enough, that u is a solution of Problem (P). To see this define v by

$$v(x)=\int_{E_0^\lambda}\psi_2(y)k^-(y)G_d^-(x,y)dy+\int_{E_\lambda^d}\psi_1(y)k^-(y)G_d^-(x,y)dy$$

$$-\int_{\Gamma_d}\psi_1(y)\frac{\partial G_d^-(x,y)}{\partial y_2}ds(y),\qquad x\in\mathbf{R}^2\backslash\overline{U}_d, \tag{3.15}$$

$$v(x)=\lim_{y\to x,\ y\in\mathbf{R}^2\backslash\overline{U}_d}v(y),\qquad x\in\Gamma_d. \tag{3.16}$$

Then, comparing (3.10) and (3.12), $\psi_1=u|_{\overline{E}_\lambda^B}$, and comparing (3.11) and (3.15), $\psi_2=v|_{\overline{E}_0^\lambda}$. Also, by Lemmas 3.1 and 3.2 applied to (3.12) and (3.15), $\psi_2=u$ on Γ_c and $\psi_1=v$ on Γ_d. Thus $u=v$ on Γ_c and Γ_d. Define $w=u-v$. Then it is clear from Lemmas 3.1 and 3.2 again, together with the above results, that (i) w is bounded in E_c^d and $w\in C(\overline{E}_c^d)\cap C^1(E_c^d)$; (ii) $\Delta w+\hat{k}^2w=0$ in E_c^d, where $\hat{k}(x)=k_-$, $x\in E_\lambda^d$, $=k_+$, $x\in E_c^\lambda$; (iii) $w=0$ on Γ_c and Γ_d. Now, consider the following eigenvalue problem: find $z\in C^1[c,d]\cap H^2(c,d)$, $\Lambda\in\mathbf{R}$, such that $-z''-qz=\Lambda z$ in (c,d) and $z(c)=z(d)=0$, where $q(x_2)=k_-$, $\lambda<x_2<d$, $=k_+$, $c<x_2<\lambda$. Provided this problem has only positive eigenvalues $\Lambda>0$, and this is the case if $(d-\lambda)k_-=(\lambda-c)k_+<\pi/2$, i.e., provided $2k_+k_-(d-c)<\pi(k_++k_-)$, then an elementary separation of variables argument establishes that $w\equiv0$ in E_c^d and hence $u\equiv v$ in E_c^d. It is now easy to see, by

further applications of Lemmas 3.1 and 3.2, that u, defined by (3.12)–(3.14), satisfies Problem (P). Thus we have the following equivalence theorem between Problem (P) and the integral equation problem (3.10) and (3.11).

THEOREM 3.2. *If $u \in C(\mathbf{R}^2)$ is a solution of Problem* (P), *then $\psi_1 := u|_{\overline{E}_\lambda^B}$ and $\psi_2 := u|_{\overline{E}_0^\lambda}$ satisfy the integral equations* (3.10) *and* (3.11). *Conversely, suppose that $\psi_1 \in BC(\overline{E}_\lambda^B)$ and $\psi_2 \in BC(\overline{E}_0^\lambda)$ satisfy the integral equations* (3.10) *and* (3.11) *and define u by* (3.12)–(3.14). *Then, provided $(d-\lambda)k_- = (\lambda - c)k_+$ and $2k_+k_-(d-c) < \pi(k_+ + k_-)$, u satisfies Problem* (P).

Remark 3.2. Let $(d-\lambda)k_- = (\lambda - c)k_+$ and $2k_+k_-(d-c) < \pi(k_+ + k_-)$. Then from Theorem 3.2 it follows that in order to prove the existence of a solution to Problem (P), it is enough to show that the pair of integral equations (3.10) and (3.11) has a solution. This will be done in section 6.

4. A basic inequality. In this section a basic inequality satisfied by solutions of (2.3) is established, which is a key step in the proof of the uniqueness theorem.

Suppose that $u \in C(\mathbf{R}^2)$ satisfies (2.3). Then, by Remark 2.3, $u \in C^1(\mathbf{R}^2) \cap H^2_{\mathrm{loc}}(\mathbf{R}^2)$. Let $\eta < c < d < \rho$ with η, ρ being as defined in assumptions (A3) or (A5) and define, for $t \in \mathbf{R}$ and $A > 0$,

$$J_A(t) = \Im \int_{\Gamma_t(A)} \overline{u}\partial_2 u ds, \qquad L_A(t) = \Re \int_{\Gamma_t(A)} \overline{u}\partial_2 u ds, \tag{4.1}$$

$$I_A^{\pm}(t) = \int_{\Gamma_t(A)} \{|\partial_2 u|^2 - |\partial_1 u|^2 + k_\pm^2 |u|^2\} ds, \tag{4.2}$$

$$K_A = \int_{E_\eta^B(A)} |u|^2 |k^2 - k_+^2| dx + \int_{E_0^\rho(A)} |u|^2 |k^2 - k_-^2| dx \tag{4.3}$$

$$+ \int_{\Gamma_c(A)} |u|^2 ds + \int_{\Gamma_d(A)} |u|^2 ds. \tag{4.4}$$

Let $a < 0 < B < b$, and for $t \in \mathbf{R}$, let $\gamma(t) = \{(t, x_2) | a \le x_2 \le b\}$.

THEOREM 4.1. *Assume that* (A3) *holds or that both* (A4) *and* (A5) *hold. Then, for some nonnegative constants C_j, $j = 1, 2, 3$, there holds*

$$\begin{aligned} K_A \le{}& C_1[(b-\beta)I_A^+(b) - (a-\beta)I_A^-(a)] + C_1[L_A(b) - L_A(a)] \\ &+ C_2[J_A(a) - J_A(b)] + C_3 R_1(A) + C_1 R_2(A), \end{aligned} \tag{4.5}$$

for all $A > 0$, where

$$R_1(A) = \left[\int_{\gamma(A)} + \int_{\gamma(-A)}\right] |\overline{u}\partial_1 u| ds$$

and

$$R_2(A) = \Re \left[\int_{\gamma(A)} - \int_{\gamma(-A)}\right] [2(x_2 - \beta)\partial_2 \overline{u}\partial_1 u + \overline{u}\partial_1 u] ds.$$

Proof. First we will deduce the inequality (4.5) in the case that (A3) holds.

Apply Green's first theorem to u and $\overline{u}$ in $E_a^b(A)$ and take the imaginary part of the result thus obtained to get that since $\Im(k^2(x)) = 0$ for $x_2 > B$ and $x_2 < 0$,

$$\int_{E_B(A)} \Im(k^2)|u|^2 dx + J_A(b) - J_A(a) \leq R_1(A). \tag{4.6}$$

Let $\theta \in C^2(\mathbf{R})$ be such that $0 \leq \theta(t) \leq 1$ for $t \in \mathbf{R}$, $\theta(t) = 1$ for $c \leq t \leq d$, and $\theta(t) = 0$ for $t \geq \rho$ and $t \leq \eta$. Then, by applying Green's first theorem to u and $\theta(x_2)\overline{u}$ in $E_B(A)$ and taking the real part of the result thus obtained, we obtain on integrating by parts that

$$\int_{E_B(A)} \theta(x_2)|\nabla u|^2 dx \leq \int_{E_B(A)} \left[\Re(k^2)\theta(x_2) + \frac{1}{2}\theta''(x_2)\right]|u|^2 dx + R_1(A),$$

which together with the definition of θ implies that

$$\int_{E_c^d(A)} |\nabla u|^2 dx \leq (\|k\|_\infty^2 + C) \int_{E_\eta^\rho(A)} |u|^2 dx + R_1(A), \tag{4.7}$$

for some constant $C > 0$ depending only on the choice of θ.

Now, for any $r, t \in \mathbf{R}$,

$$u((x_1, r)) - u((x_1, t)) = \int_t^r \partial_2 u(x) dx_2, \qquad x_1 \in \mathbf{R}, \tag{4.8}$$

so that using the Cauchy–Schwarz inequality,

$$|u((x_1, t))|^2 \leq 2|u((x_1, r))|^2 + 2(r - t) \int_t^r |\partial_2 u(x)|^2 dx_2, \qquad x_1 \in \mathbf{R}. \tag{4.9}$$

From (4.9) it follows that, for $R < T$, $r, t \in [R, T]$,

$$\int_{\Gamma_t(A)} |u|^2 ds \leq 2 \int_{\Gamma_r(A)} |u|^2 ds + 2(T - R) \int_{E_R^T(A)} |\partial_2 u|^2 dx \tag{4.10}$$

and hence that

$$(T - R) \int_{\Gamma_t(A)} |u|^2 ds \leq 2 \int_{E_R^T(A)} |u|^2 dx + 2(T - R)^2 \int_{E_R^T(A)} |\partial_2 u|^2 dx. \tag{4.11}$$

Thus, assuming that (A3) holds, the required inequality (4.5), with $C_1 = 0$, follows from (4.6), (4.7), and (4.11) with $R = c$, $T = d$, $t = c, d$.

Suppose now that (A4) and (A5) hold. Multiplying (2.3) by $2(x_2 - \beta)\partial_2\overline{u} + \overline{u}$, integrating over $E_a^b(A)$, and taking the real part, we obtain on noting that $\Im(k^2(x)) = 0$ for $x_2 > B$ and $x_2 < 0$,

$$2 \int_{E_a^b(A)} |\partial_2 u|^2 dx = \Re \int_{E_a^b(A)} \{2\nabla \cdot [(x_2 - \beta)\partial_2\overline{u}\nabla u] - \partial_2[(x_2 - \beta)|\nabla u|^2] + \nabla \cdot (\overline{u}\nabla u)\} dx$$

$$
\begin{aligned}
&+\int\limits_{E_a^b(A)} \Re(k^2)\partial_2[(x_2-\beta)|u|^2]dx + 2\int\limits_{E_B(A)} (x_2-\beta)\Im(k^2)\Im\{\overline{u}\partial_2 u\}dx \\
&= (b-\beta)\int\limits_{\Gamma_b(A)} (|\partial_2 u|^2 - |\partial_1 u|^2)ds - (a-\beta)\int\limits_{\Gamma_a(A)} (|\partial_2 u|^2 - |\partial_1 u|^2)ds \\
&+L_A(b) - L_A(a) + R_2(A) + \int\limits_{E_a^b(A)} \Re(k^2)\partial_2[(x_2-\beta)|u|^2]dx \\
&+2\int\limits_{E_B(A)} (x_2-\beta)\Im(k^2)\Im\{\overline{u}\partial_2 u\}dx. \qquad (4.12)
\end{aligned}
$$

Now, if $\Re(k^2) \in C^1(\mathbf{R}^2)$, then from (A4) we have that $(x_2-\beta)\partial_2(\Re(k^2)) \geq 0$, and integrating by parts, we obtain that

$$
\begin{aligned}
\int\limits_{E_a^b(A)} \Re(k^2)\partial_2[(x_2-\beta)|u|^2]dx &\leq (b-\beta)k_+^2 \int\limits_{\Gamma_b(A)} |u|^2 ds - (a-\beta)k_-^2 \int\limits_{\Gamma_a(A)} |u|^2 ds \\
&= \int\limits_{E_a^b(A)} \tilde{k}^2\partial_2[(x_2-\beta)|u|^2]dx. \qquad (4.13)
\end{aligned}
$$

Thus

$$
G_A \equiv \int\limits_{E_a^b(A)} [\tilde{k}^2 - \Re(k^2)]\partial_2[(x_2-\beta)|u|^2]dx \geq 0. \qquad (4.14)
$$

In the general case where $k \in L_\infty(\mathbf{R}^2)$, let $\phi(x) = \Re(k^2(x))$, $\psi(x) = (x_2-\beta)|u(x)|^2$, and for $h \in \mathbf{R}$, let $\phi_h(x) = \phi(x+he_2)$, $\psi_h(x) = \psi(x+he_2)$. Then, since $\phi(\psi_h - \psi) + \phi_h(\psi_h - \psi) = 2(\phi_h\psi_h - \phi\psi) - (\phi_h - \phi)(\psi + \psi_h)$, we have that for sufficiently small $h > 0$,

$$
\begin{aligned}
&\int\limits_{E_a^b(A)} \phi(\psi_h - \psi)dx + \int\limits_{E_{a+h}^{b+h}(A)} \phi(\psi - \psi_{-h})dx \\
&= 2\int\limits_{E_b^{b+h}(A)} \phi\psi dx - 2\int\limits_{E_a^{a+h}(A)} \phi\psi dx - \int\limits_{E_a^b(A)} (\phi_h - \phi)(\psi + \psi_h)dx. \qquad (4.15)
\end{aligned}
$$

By using (A4) the last term on the right-hand side of (4.15) can be estimated as follows. First, in the cases where $\beta \leq a$ and $\beta \geq b+h$, it is easy to see that

$$
I \equiv -\int\limits_{E_a^b(A)} (\phi_h - \phi)(\psi + \psi_h)dx \leq 0,
$$

while if $a < \beta < b+h$, then

$$
I \leq 2||k||_\infty^2 \int\limits_{E_{\beta-h}^{\beta}(A)} |\psi + \psi_h|dx \equiv I_h.
$$

Therefore, $I \le I_h$ in any case, and thus it follows from (4.15) that

$$\int\limits_{E_a^b(A)} \phi(\psi_h - \psi)dx + \int\limits_{E_{a+h}^{b+h}(A)} \phi(\psi - \psi_{-h})dx$$

$$\le 2k_+^2 \int\limits_{E_b^{b+h}(A)} \psi dx - 2k_-^2 \int\limits_{E_a^{a+h}(A)} \psi dx + I_h \tag{4.16}$$

on using (A2). Since $\psi \in C^1(\mathbf{R}^2)$ and $\psi = 0$ on Γ_β, dividing (4.16) by $2h$ and taking the limit $h \to 0$ we obtain that (4.13) and (4.14) hold in the general case.

It follows from (4.12) that

$$2\int\limits_{E_a^b(A)} |\partial_2 u|^2 dx + G_A = (b-\beta)I_A^+(b) - (a-\beta)I_A^-(a) + L_A(b) - L_A(a) + R_2(A)$$

$$+2\int\limits_{E_B(A)} (x_2 - \beta)\Im(k^2)\Im\{\overline{u}\partial_2 u\}dx. \tag{4.17}$$

Since $0 \le \Im(k^2) \le \|k\|_\infty^2$, the Cauchy–Schwarz inequality yields that

$$2\int\limits_{E_B(A)} (x_2 - \beta)\Im(k^2)\Im\{\overline{u}\partial_2 u\}dx \le \int\limits_{E_B(A)} |\partial_2 u|^2 dx + (B + |\beta|)^2\|k\|_\infty^2 \int\limits_{E_B(A)} \Im(k^2)|u|^2 dx.$$

(4.18)

Thus, it follows from (4.17), (4.18), and (4.6) that

$$\int\limits_{E_a^b(A)} |\partial_2 u|^2 dx + G_A \le (b-\beta)I_A^+(b) - (a-\beta)I_A^-(a) + L_A(b) - L_A(a) + R_2(A)$$

$$+(B + |\beta|)^2\|k\|_\infty^2[J_A(a) - J_A(b) + R_1(A)] \equiv F_A. \tag{4.19}$$

Now, from (4.19) and the fact that $G_A \ge 0$, it is seen that

$$\int\limits_{E_a^b(A)} |\partial_2 u|^2 dx \le F_A. \tag{4.20}$$

On the other hand, since $2|(x_2 - \beta)\Re(\overline{u}\partial_2 u)| \le |u|^2/2 + 2(b + |\beta|)^2|\partial_2 u|^2$ in E_a^b, we have

$$\partial_2[(x_2 - \beta)|u|^2] = |u|^2 + 2(x_2 - \beta)\Re(\overline{u}\partial_2 u) \ge |u|^2/2 - 2(b + |\beta|)^2|\partial_2 u|^2,$$

for $x \in E_a^b$, so that on noting that $\tilde{k}^2 \ge \Re(k^2)$ by (A4),

$$G_A \ge \frac{1}{2}\int\limits_{E_a^b(A)} [\tilde{k}^2 - \Re(k^2)]|u|^2 dx - 4(b + |\beta|)^2||k||_\infty^2 \int\limits_{E_a^b(A)} |\partial_2 u|^2 dx.$$

This, together with (4.19) and (4.20), implies that

$$\int\limits_{E_a^b(A)} [\tilde{k}^2 - \Re(k^2)]|u|^2 dx \le 2[1 + 4(b + |\beta|)^2||k||_\infty^2]F_A. \tag{4.21}$$

We now make use of (4.20) and (4.21) to derive the required inequality (4.5). First, using (4.9) and the fact that $\Re(k^2) \le \tilde{k}^2$ by (A4), we obtain (cf. (4.11)) that

$$\int_{E_c^d(A)} [\tilde{k}^2 - \Re(k^2)]|u((x_1, c))|^2 dx$$

$$\le 2 \int_{E_c^d(A)} [\tilde{k}^2 - \Re(k^2)]|u|^2 dx + 4(d-c)^2 ||k||_\infty^2 \int_{E_c^d(A)} |\partial_2 u|^2 dx. \tag{4.22}$$

Using (A4) and (A5) yields that

$$(d-c)\lambda_3 \int_{\Gamma_c(A)} |u|^2 ds \le 2 \int_{E_a^b(A)} [\tilde{k}^2 - \Re(k^2)]|u|^2 dx + 4B^2 ||k||_\infty^2 \int_{E_a^b(A)} |\partial_2 u|^2 dx. \tag{4.23}$$

From (4.10) with $R = a$, $T = b$, $r = c$, we obtain that

$$\int_{E_a^b(A)} |u|^2 dx \le 2(b-a) \int_{\Gamma_c(A)} |u|^2 ds + 2(b-a)^2 \int_{E_a^b(A)} |\partial_2 u|^2 dx. \tag{4.24}$$

Thus, utilizing (4.20) and (4.21) together with (4.23), (4.24), and (4.11), with $R = a$, $T = b$, $t = d$, it follows that K_A is bounded by a multiple of F_A. Thus the required result (4.5) holds with $C_2 = C_3$. □

5. Uniqueness of solution. In this section we establish the following uniqueness theorem for Problem (P).

THEOREM 5.1. *If* (A3) *holds or both* (A4) *and* (A5) *hold, then Problem* (P) *has at most one solution.*

We prove this theorem by showing that the homogeneous version of Problem (P) has only the trivial solution. Since guided waves are solutions of the homogeneous problem (see Definition A.1 and Theorem A.1 in the appendix), we have immediately the following corollary.

COROLLARY 5.1. *If* (A3) *holds or both* (A4) *and* (A5) *hold, then there are no guided wave solutions to the homogeneous problem.*

In the proof of Theorem 5.1 we utilize the following two lemmas, the first of which is a special case of Lemma A in [8].

LEMMA 5.1. *Suppose that* $F \in L_\infty(\mathbf{R})$ *and that, for some nonnegative constants* $C, \epsilon,$ *and* A_0,

$$\int_{-A}^{A} |F(t)|^2 dt \le C \int_{\mathbf{R}\setminus[-A,A]} G_A^2(t) dt + C \int_{-A}^{A} (G_\infty(t) - G_A(t)) G_\infty(t) dt + \epsilon, \quad A > A_0,$$

where, for $A_0 < A \le +\infty$,

$$G_A(s) = \int_{-A}^{A} (1 + |s-t|)^{-3/2} |F(t)| dt, \qquad s \in \mathbf{R}.$$

Then $F \in L_2(\mathbf{R})$ *and*

$$\int_{-\infty}^{+\infty} |F(t)|^2 dt \le \epsilon.$$

LEMMA 5.2. *If $\phi_+ \in L_2(\Gamma_H) \cap L_\infty(\Gamma_H)$, $\phi_- \in L_2(\Gamma_h) \cap L_\infty(\Gamma_h)$, and $v_\pm$ are defined by* (2.4) *and* (2.5), *respectively, then the restrictions of v_+, $\partial_1 v_+$, and $\partial_2 v_+$ to Γ_b are in $L_2(\Gamma_b) \cap BC(\Gamma_b)$ for $b > H$; the restrictions of v_-, $\partial_1 v_-$, and $\partial_2 v_-$ to Γ_a are in $L_2(\Gamma_a) \cap BC(\Gamma_a)$ for $a < h$; and*

$$\Im \int_{\Gamma_b} \overline{v}_+ \partial_2 v_+ ds \geq 0, \qquad \Re \int_{\Gamma_b} \overline{v}_+ \partial_2 v_+ ds \leq 0, \tag{5.1}$$

$$\int_{\Gamma_b} [|\partial_2 v_+|^2 - |\partial_1 v_+|^2 + k_+^2 |v_+|^2] ds \leq 2k_+ \Im \int_{\Gamma_b} \overline{v}_+ \partial_2 v_+ ds, \tag{5.2}$$

and

$$\Im \int_{\Gamma_a} \overline{v}_- \partial_2 v_- ds \leq 0, \qquad \Re \int_{\Gamma_a} \overline{v}_- \partial_2 v_- ds \geq 0, \tag{5.3}$$

$$\int_{\Gamma_a} [|\partial_2 v_-|^2 - |\partial_1 v_-|^2 + k_-^2 |v_-|^2] ds \leq -2k_- \Im \int_{\Gamma_a} \overline{v}_- \partial_2 v_- ds. \tag{5.4}$$

The statements in this lemma concerning v_+ were proved as in Lemma 6.1 in [7]. The statements regarding v_- follow from Remark 2.1.

Proof of Theorem 5.1. As in section 4, let $\eta < c < d < \rho$ and $a < 0$, $b > B$. Also for convenience choose a and b so that $b - \beta = \beta - a \equiv \omega > 0$.

Suppose that u_1 and u_2 are solutions of Problem (P). Then, by Remark 2.3, $u = u_1 - u_2 \in C^1(\mathbf{R}^2)$ and satisfies (2.3), the bound (2.8), the UPRC, and the DPRC. Also, by Theorem 3.1,

$$u(x) = \int_{E_c^B} u(y) k^+(y) G_c^+(x,y) dy + \int_{\Gamma_c} u(y) \frac{\partial G_c^+(x,y)}{\partial y_2} ds(y), \qquad x \in U_c, \tag{5.5}$$

$$u(x) = \int_{E_0^d} u(y) k^-(y) G_d^-(x,y) dy - \int_{\Gamma_d} u(y) \frac{\partial G_d^-(x,y)}{\partial y_2} ds(y), \qquad x \in \mathbf{R}^2 \backslash \overline{U}_d, \tag{5.6}$$

and by Theorem 4.1, for some constants $C_j \geq 0$, $j = 1, 2, 3$,

$$\begin{aligned} K_A \leq\ & C_1[\omega I_A^+(b) + \omega I_A^-(a) + L_A(b) - L_A(a) + R_2(A)] \\ & + C_2[J_A(a) - J_A(b)] + C_3 R_1(A), \end{aligned} \tag{5.7}$$

where J_A, $I_A^\pm$, L_A, and K_A are given by (4.1) and (4.2). Clearly, for $j = 1, 2$,

$$R_j(A) = O(1) \text{ as } A \to \infty, \tag{5.8}$$

and by (4.6),

$$J_A(b) - J_A(a) \leq R_1(A). \tag{5.9}$$

Now to make use of Lemma 5.1 and the bound (5.7), we define

$$v(x) = \int_{E_c^B(A)} u(y) k^+(y) G_c^+(x,y) dy + \int_{\Gamma_c(A)} u(y) \frac{\partial G_c^+(x,y)}{\partial y_2} ds(y), \quad x \in U_c, \tag{5.10}$$

$$v(x) = \int_{E_0^d(A)} u(y) k^-(y) G_d^-(x,y) dy - \int_{\Gamma_d(A)} u(y) \frac{\partial G_d^-(x,y)}{\partial y_2} ds(y), \quad x \in \mathbf{R}^2 \backslash \overline{U}_c. \tag{5.11}$$

Then, by (3.1)–(3.2), $v|_{\Gamma_B} \in L_2(\Gamma_B) \cap BC(\Gamma_B)$ and $v|_{\Gamma_0} \in L_2(\Gamma_0) \cap BC(\Gamma_0)$. Moreover, by Lemmas 3.1 and 3.2 and the equivalence of Lemmas 2.1(i)–(ii) and 2.2, v satisfies (2.4), with $h = B$ and $\phi_+ = v|_{\Gamma_B}$, and satisfies (2.5) with $h = 0$ and $\phi_- = v|_{\Gamma_0}$. For $t \in \mathbf{R}$ set,

$$J_A'(t) = \Im \int_{\Gamma_t(A)} \overline{v}\partial_2 v ds, \qquad J_A''(t) = \Im \int_{\Gamma_t} \overline{v}\partial_2 v ds,$$

$$I_A^{\pm\prime}(t) = \int_{\Gamma_t(A)} \{|\partial_2 v|^2 - |\partial_1 v|^2 + k_\pm^2 |v|^2\} ds, \quad I_A^{\pm\prime\prime} = \int_{\Gamma_t} \{|\partial_2 v|^2 - |\partial_1 v|^2 + k_\pm^2 |v|^2\} ds,$$

$$L_A'(t) = \Re \int_{\Gamma_t(A)} \overline{v}\partial_2 v ds, \qquad L_A''(t) = \Re \int_{\Gamma_t} \overline{v}\partial_2 v ds.$$

Then, by Lemma 5.2,

$$\begin{aligned} &J_A''(b) \geq 0, \qquad L_A''(b) \leq 0, \qquad I_A^{+\prime\prime}(b) \leq 2k_+ J_A''(b), \\ &J_A''(a) \leq 0, \qquad L_A''(a) \geq 0, \qquad I_A^{-\prime\prime}(a) \leq -2k_- J_A''(a). \end{aligned}$$

Hence, by the preceding and (5.7),

$$\begin{aligned} K_A \leq C_1 \{ &\omega [\ I_A^+(b) - I_A^{+''}(b)] + \omega [I_A^-(a) - I_A^{-''}(a)] + [L_A(b) - L_A''(b)] + [L_A''(a) - L_A(a)]\} \\ &+ [C_2 + 2C_1\omega(k_+ + k_-)]\{[J_A''(b) - J_A(b)] + [J_A(a) - J_A''(a)]\} \\ &+ [C_3 + 2C_1\omega(k_+ + k_-)] R_1(A) + C_1 R_2(A). \end{aligned} \tag{5.12}$$

Now note that

$$K_A = \int_{-A}^{A} |w(x_1)|^2 dx_1,$$

where

$$\begin{aligned} w(x_1) = \Big\{ &\int_\eta^B |u(x)|^2 |k^2(x) - k_+^2| dx_2 + \int_0^\rho |u(x)|^2 |k^2(x) - k_-^2| dx_2 \\ &+ |u(x_1, c)|^2 + |u(x_1, d)|^2 \Big\}^{1/2}, \qquad x_1 \in \mathbf{R}, \end{aligned}$$

and note that by (3.3) and the Cauchy–Schwarz inequality, for $x \in \Gamma_a,\ \Gamma_b$,

$$\begin{aligned} &|v(x)|, \qquad |\nabla v(x)| \leq C W_A(x_1), \\ &|u(x) - v(x)|, \qquad |\nabla u(x) - \nabla v(x)| \leq C(W_\infty(x_1) - W_A(x_1)), \end{aligned}$$

where C is a constant independent of A and, for $0 \leq A \leq +\infty$,

$$W_A(x_1) = \int_{-A}^{A} (1 + |x_1 - y_1|)^{-3/2} w(y_1) dy_1, \qquad x_1 \in \mathbf{R}.$$

It follows that

$$\begin{aligned} |I_A^{\pm\prime}(t) - I_A^{\pm\prime\prime}(t)|, \qquad &|J_A'(t) - J_A''(t)|, \qquad |L_A'(t) - L_A''(t)| \\ &\leq C \int_{\mathbf{R}\setminus[-A,A]} (W_A(x_1))^2 dx_1, \qquad (t = a, b), \end{aligned}$$

where C is a constant independent of A, and that

$$|I_A^{\pm}(t) - I_A^{\pm\prime}(t)|, \qquad |J_A(t) - J_A'(t)|, \qquad |L_A(t) - L_A'(t)|$$
$$\leq C \int_{-A}^{A} (W_\infty(x_1) - W_A(x_1)) W_\infty(x_1) dx_1, \qquad (t = a, b),$$

so that, from (5.12) for some constant $C_0 > 0$ and all $A > 0$,

$$K_A \leq C_0 \Big\{ \int_{\mathbf{R}\setminus[-A,A]} W_A^2(x_1) dx_1 + \int_{-A}^{A} (W_\infty(x_1) - W_A(x_1)) W_\infty(x_1) dx_1$$
$$+ |R_1(A)| + |R_2(A)| \Big\}. \tag{5.13}$$

Applying Lemma 5.1 to (5.13) we obtain that $w \in L_2(\mathbf{R})$, i.e., $u \in L_2(E_B) \cap L_2(\Gamma_c) \cap L_2(\Gamma_d)$ and, for all $A_0 > 0$,

$$\int_{E_B} |u|^2 dx + \int_{\Gamma_c} |u|^2 ds + \int_{\Gamma_d} |u|^2 ds = \int_{-\infty}^{+\infty} |w|^2$$
$$\leq C_0 \sup_{A > A_0} (|R_1(A)| + |R_2(A)|). \tag{5.14}$$

Since $u \in L_2(E_B) \cap L_2(\Gamma_c) \cap L_2(\Gamma_d)$, it follows from (5.5), (5.6), the bounds (3.1) and (3.2), and applications of Young's theorem that $u \in L_2(E_a^b)$ for any $a, b \in \mathbf{R}$ with $a < b$. Also, since $\nabla u \in BC(E_a^b)$ so that u is uniformly continuous in $\overline{E_a^b}$, it follows that $u(x) \to 0$ as $x_1 \to \infty$ uniformly in x_2 for $a \leq x_2 \leq b$ for any real numbers $a < b$. Also, noting Lemma 2.2, it follows that $R_j(A) \to 0$ as $A \to \infty$, $j = 1, 2$, and thus, from (5.14), that $u = 0$ in E_B and on $\Gamma_c \cup \Gamma_d$; and hence, from (5.5) and (5.6), that $u \equiv 0$ in $\mathbf{R}^2$. □

6. Existence of solution. In this section existence of a solution for Problem (P) will be established by making use of general results on the solvability of the system of second-kind integral equations

$$\psi_i = \phi_i + \sum_{j=1}^{N} K_{ij} \psi_j, \qquad i = 1, \ldots, N, \tag{6.1}$$

in which $\phi_i \in Y_i := BC(\overline{\Omega}_i)$ is assumed known, $\psi_i \in Y_i$ is to be determined, and $K_{ij} : Y_j \to Y_i$ is the integral operator defined by

$$K_{ij}\psi(x) = \int_{\Omega_j} k_{ij}(x, y) \psi(y) d\mu_j(y), \qquad x \in \overline{\Omega}_i, \tag{6.2}$$

$i, j = 1, \ldots, N$. Here Ω_j is an open subset of $\mathbf{R}^{n_j}$ ($n_j \geq 1$) and $d\mu_j$ is n_j-dimensional Lebesgue measure. The function $k_{ij} : \overline{\Omega}_i \times \Omega_j \to \mathbf{C}$ is assumed to take the form, for some $M \in \mathbf{N}$,

$$k_{ij}(x, y) = \sum_{m=1}^{M} k_{ij}^{(m)}(x, y) z_j^{(m)}(y),$$

where $z_j^{(m)} \in X_j := L_\infty(\Omega_j)$ and $k_{ij}^{(m)}(x,\cdot) \in L_1(\Omega_j)$ for every $x \in \overline{\Omega}_i$ $(i,j = 1,\dots,N,\ m = 1,\dots,M)$. We assume that the following conditions on $k_{ij}^{(m)}$ and Ω_j hold:

(C.1) $\sup_{x\in\overline{\Omega}_i} \int_{\Omega_j} |k_{ij}^{(m)}(x,y)|d\mu_j(y) < \infty$ and, for all $x \in \overline{\Omega}_i$,

$$\int_{\Omega_j} |k_{ij}^{(m)}(x,y) - k_{ij}^{(m)}(x',y)|d\mu_j(y) \to 0$$

as $x' \to x$ with $x' \in \overline{\Omega}_i$ $(i,j = 1,\dots,N,\ m = 1,\dots,M)$.

(C.2) For some $n_0 \le \min_j n_j$ and $i = 1,\dots,N$, there exists $a_j^{(i)} \in \mathbf{R}^{n_i}$, $j = 1,\dots,n_0$, and a bounded set $\omega_i \subset \Omega_i$ such that

(i) $\Omega_i = \bigcup_{P\in\mathbf{Z}^{n_0}} \omega_i^{(P)}$, where $\omega_i^{(P)} := \omega_i + \sum_{j=1}^{n_0} a_j^{(i)} p_j$, for $P = (p_1,\dots,p_{n_0}) \in \mathbf{Z}^{n_0}$;

(ii) $\omega_i^{(Q)} \cap \omega_i^{(P)} = \emptyset$ for $Q, P \in \mathbf{Z}^{n_0}$, $Q \ne P$;

(iii) $k_{ij}^{(m)}(x + a_l^{(i)}, y + a_l^{(j)}) = k_{ij}^{(m)}(x,y)$, $x \in \overline{\Omega}_i$, $y \in \Omega_j$, $i,j = 1,\dots,N$, $l = 1,\dots,n_0$,

$m = 1,\dots,M$.

Let X and Y denote the product spaces $X := \prod_{j=1}^N X_j$ and $Y := \prod_{j=1}^N Y_j \subset X$. Let $\phi = (\phi_1,\dots,\phi_N)^t$, $\psi = (\psi_1,\dots,\psi_N)^t \in Y$, where $(\cdot,\dots,\cdot)^t$ denotes the transpose of $(\cdot,\dots,\cdot)$. For $m = 1,\dots,M$, define the matrix operator $K^{(m)}$ on X by

$$K^{(m)} = \begin{pmatrix} K_{11}^{(m)} & \cdots & K_{1N}^{(m)} \\ & \ddots & \\ K_{N1}^{(m)} & \cdots & K_{NN}^{(m)} \end{pmatrix}, \tag{6.3}$$

where $K_{ij}^{(m)} : X_j \to X_i$ is the integral operator defined by (6.2) with K_{ij}, k_{ij} replaced by $K_{ij}^{(m)}$, $k_{ij}^{(m)}$. For $z = (z_1,\dots,z_N)^t \in X$ define $\hat{z}$ by

$$\hat{z} = \begin{pmatrix} z_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & z_N \end{pmatrix},$$

and for $m = 1,\dots,M$ and $z \in X$, define $K_z^{(m)} : Y \to Y$ by

$$K_z^{(m)}\psi = K^{(m)}(\hat{z}\psi), \qquad \psi \in Y. \tag{6.4}$$

For $w = (w^{(1)},\dots,w^{(M)}) \in X^M$ let K_w denote the matrix integral operator

$$K_w = \sum_{m=1}^M K_{w^{(m)}}^{(m)}. \tag{6.5}$$

Then (6.1) can be abbreviated as

$$\psi = \phi + K_z\psi, \tag{6.6}$$

where $z = (z^{(1)},\dots,z^{(M)})$ and $z^{(m)} = (z_1^{(m)},\dots,z_N^{(m)})^t$, $m = 1,\dots,M$.

For $j = 1, \dots, N$, $i = 1, \dots, n_0$, define the translation operator $T_{a_j^{(i)}} : X_j \to X_j$ by

$$T_{a_j^{(i)}} \psi(x) = \psi(x - a_j^{(i)}), \qquad x \in \Omega_j,$$

and for $a = (a_1, \dots, a_N) \in \tau := \{(a_l^{(1)}, \dots, a_l^{(N)})^t | l = 1, \dots, n_0\}$, define the matrix operator $T_a : X \to X$ by

$$T_a = \begin{pmatrix} T_{a_1} & & 0 \\ & \ddots & \\ 0 & & T_{a_N} \end{pmatrix}.$$

Then, by (C.2) (iii), $T_a K^{(m)} = K^{(m)} T_a$, $a \in \tau$, $m = 1, \dots, M$. Let $B(Y)$ denote the Banach space of bounded linear operators on Y and I the identity matrix operator on Y. The following results have been proved in [10], extending the results of [9] for single integral equations to systems of integral equations.

THEOREM 6.1. *Suppose that* (C.1) *and* (C.2) *are satisfied, that* $W \subset X^M$ *is weak* sequentially compact, that* $T_a W := \{(T_a z^{(1)}, \dots, T_a z^{(M)}) | (z^{(1)}, \dots, z^{(M)}) \in W\} = W$, $a \in \tau$, *and that* $I - K_z$ *is injective for all* $z \in W$. *Then* $(I - K_z)^{-1}$ *exists as an operator on the range space* $(I - K_z)Y$ *for all* $z \in W$ *and*

$$\sup_{z \in W} \|(I - K_z)^{-1}\| < \infty.$$

Also, if for every $z \in W$ *there exists a sequence* $(z_j) \subset W$ *such that* (z_j) *converges weak* to* z *in* X *and*

$$\text{for all } j, \qquad I - K_{z_j} \quad \text{injective} \Longrightarrow I - K_{z_j} \text{ surjective},$$

then $I - K_z$ *is surjective also for each* $z \in W$ *so that* $(I - K_z)^{-1} \in B(Y)$.

THEOREM 6.2. *If* (C.1) *and* (C.2) *are satisfied,* $z = ((z_1^{(1)}, \dots, z_N^{(1)})^t, \dots, (z_1^{(M)}, \dots, z_N^{(M)})^t) \in X^M$, *and for some constants* $\lambda_j^{(m)} \in \mathbf{C}$, $j = 1, \dots, N$, $m = 1, \dots, M$, *it holds that*

$$\text{ess sup}_{|x| \geq A, x \in \Omega_j} |z_j^{(m)}(x) - \lambda_j^{(m)}| \to 0$$

as $A \to \infty$, *then*

$$I - K_\lambda, \qquad I - K_z \text{ injective } \Rightarrow I - K_z \text{ surjective}, \qquad (I - K_z)^{-1} \in B(Y),$$

where $\lambda = ((\lambda_1^{(1)}, \dots, \lambda_N^{(1)})^t, \dots, (\lambda_1^{(M)}, \dots, \lambda_N^{(M)})^t)$.

To apply Theorems 6.1 and 6.2 to show the existence of a solution to Problem (P), we choose λ, c, and d so that $0 \leq c < d \leq B$, $k_-(d - \lambda) = k_+(\lambda - c)$, and $2(d - c)k_+k_- < \pi(k_+ + k_-)$. It then follows from Theorem 3.2 that Problem (P) and the integral equation problems (3.10) and (3.11) are equivalent.

Let $N = 4$, $n_0 = 1$, $\Omega_1 = E_\lambda^B \subset \mathbf{R}^2$, $\Omega_2 = E_0^\lambda \subset \mathbf{R}^2$, $\Omega_3 = \Omega_4 = \mathbf{R}$, $\omega_1 = \{x \in \Omega_1 | 0 \leq x_1 < B, \lambda < x_2 < B\}$, $\omega_2 = \{x \in \Omega_2 | 0 \leq x_1 < B, 0 < x_2 < \lambda\}$, $\omega_3 = \omega_4 = [0, B)$, $a_1^{(1)} = a_1^{(2)} = (B, 0)$, and $a_1^{(3)} = a_1^{(4)} = B$. Define $\tilde{y} := y$ and $\hat{y} := y$ for $y \in \mathbf{R}^2$ and $\tilde{y} := (y, c)$ and $\hat{y} := (y, d)$ for $y \in \mathbf{R}$. Let $\chi(t) = 1$, $t > 0$, $= 0$, $t < 0$, and let $M = 2$, $k_{ij}^{(1)}(x, y) = G_c^+(\hat{x}, y)\chi(y_2 - c)$ for all $x \in \Omega_i$, $y \in \Omega_j$,

$\hat{x} \neq y$, $i = 1, 3, j = 1, 2$, $k_{ij}^{(1)}(x, y) = G_d^-(\tilde{x}, y)\chi(d - y_2)$ for all $x \in \Omega_i$, $y \in \Omega_j$, $\tilde{x} \neq y$, $i = 2, 4, j = 1, 2$, $k_{i4}^{(1)}(x, y) = \partial G_c^+(\hat{x}, z)/\partial z_2|_{z=\tilde{y}}$ for all $x \in \Omega_i$, $y \in \Omega_4$, $i = 1, 3$, $k_{i3}^{(1)}(x, y) = \partial G_d^-(\tilde{x}, z)/\partial z_2|_{z=\hat{y}}$ for all $x \in \Omega_i$, $y \in \Omega_3$, $i = 2, 4$, $k_{ij}^{(1)}(x, y) = 0$ for all $x \in \Omega_i$, $y \in \Omega_j$, with $i = 1, 3, j = 3$, or $i = 2, 4, j = 4$. Let $k_{ij}^{(2)} = k_{ij}^{(1)}$, $i = 2, 4$, $j = 1, 2, = 0$, otherwise. Then conditions (C.1) and (C.2) are satisfied with these choices of $k_{ij}^{(m)}$ and Ω_j $(i, j = 1, 2, 3, 4, m = 1, 2)$. Set $w_j^{(1)}(y) = k^+(y)$ for $y \in \Omega_j$, $j = 1, 2$, $w_3^{(1)}(y) = -1$, $y \in \Omega_3$, $w_4^{(1)}(y) = 1$, $y \in \Omega_4$, and set $w^{(1)} = (w_1^{(1)}, \ldots, w_4^{(1)})^t$, $w^{(2)} = (k_+^2 - k_-^2)(1, 1, 0, 0)^t$. Then the integral equations (3.10) and (3.11) can be written as the 4×4 matrix system

$$(I - K_w)\psi = \phi, \qquad \psi = (\psi_1, \ldots, \psi_4)^t, \qquad \phi = (\phi_1, \ldots, \phi_4)^t \in Y, \tag{6.7}$$

where $w = (w^{(1)}, w^{(2)})$, K_w is defined by (6.5), (6.4), and (6.3), $\phi_2 = \phi_4 = 0$, $\phi_j(y) = u^i(\hat{y}) + u^r(\hat{y})$, $y \in \overline{\Omega}_j$, $j = 1, 3$, and $\psi_3, \psi_4 \in BC(\mathbf{R})$ are defined by $\psi_3(y) = \psi_2(\hat{y})$, $\psi_4(y) = \psi_1(\tilde{y})$, $y \in \mathbf{R}$.

THEOREM 6.3. *Assume that* (A3) *holds or that* (A4) *and* (A5) *hold and that* $k_-(d - \lambda) = k_+(\lambda - c)$ *and* $2(d - c)k_+k_- < \pi(k_+ + k_-)$. *Then* $(I - K_w)^{-1} \in B(Y)$ *so that the system of integral equations* (6.7) *has a unique solution* $\psi \in Y$. *Furthermore, for any* $L > 0$, *there is a constant* $C > 0$ *depending only on* L, $k_\pm$, c, d, η, ρ, B, λ_1, *and* λ_2, *in the case that* (A3) *is satisfied, or on* L, $k_\pm$, c, d, η, ρ, B, β, *and* λ_3, *in the case that* (A4) *and* (A5) *are satisfied such that, provided* $\|k\|_\infty \leq L$, $\|(I - K_w)^{-1}\| \leq C$ *so that* $\|\psi\| \leq C\|\phi\|$.

Proof. Theorem 6.3 is proved by means of Theorems 6.1 and 6.2. To this end, suppose without loss of generality that $L > k_\pm^2 + \lambda_1$ and set

$$\begin{aligned} Q = Q_3 := \{&\mu \in L_\infty(E_B) | \Im\mu \geq 0, \Im\mu(x) \geq \lambda_1, x \in E_\eta^\rho, \Im\mu(x) \geq \lambda_2 |\mu(x) - k_+^2|, \\ &x \in E_\eta^B, \Im\mu(x) \geq \lambda_2 |\mu(x) - k_-^2|, x \in E_0^\rho, ||\mu||_\infty \leq L^2\} \end{aligned}$$

in the case that (A3) is satisfied. In the case that (A4) and (A5) are satisfied suppose without loss of generality that $L^2 > k_\pm^2 + \lambda_3$ and set

$$\begin{aligned} Q = Q_4 := \{&\mu \in L_\infty(\mathbf{R}^2) | \Im\mu \geq 0, \mu(x) = k_+^2, x \in U_B, \mu(x) = k_-^2, x \in \mathbf{R}^2 \backslash \overline{U}, \\ &||\mu||_\infty \leq L^2, \text{ ess inf}_{x \in U_\beta} \Re[\mu(x + e_2 h) - \mu(x)] \geq 0, \\ &\text{ess inf}_{x \in \mathbf{R}^2 \backslash \overline{U}_\beta} \Re[\mu(x - e_2 h) - \mu(x)] \geq 0, \ h > 0, \\ &\text{ess inf}_{x \in E_\eta^\rho} \{\tilde{k}^2(x) - \Re[\mu(x)]\} \geq \lambda_3\}. \end{aligned}$$

Define $W^{(1)} \subset X$ by

$$W^{(1)} = \{(\mu|_{\Omega_1} - k_+^2, \ \mu|_{\Omega_2} - k_+^2, \ -1, \ 1)^t | \mu \in Q\}$$

and $W \subset X^2$ by

$$W = \{(w^{(1)}, \ (k_+^2 - k_-^2, \ k_+^2 - k_-^2, \ 0, \ 0)^t) | w^{(1)} \in W^{(1)}\}.$$

Then $T_a W = W$ for $a \in \tau = \{(a_1^{(1)}, \ldots, a_1^{(4)})\}$. Also, it follows easily from Theorems 3.2 and 5.1 that $I - K_z$ is injective for all $z \in W$.

Next, we show that W is weak$*$ sequentially compact. In view of the definition of W it is sufficient to show that $W^{(1)} \subset X$ is weak$*$ sequentially compact. Further,

in view of the definition of $W^{(1)}$, it is sufficient to show that Q is weak$*$ sequentially compact, where $Q = Q_3 \subset L_\infty(E_B)$ in the case that (A3) is satisfied and $Q = Q_4 \subset L_\infty(\mathbf{R}^2)$ in the case that (A4) and (A5) are satisfied.

Since Q is bounded it follows from the Alaoglu theorem [16, p. 60] that Q is weak$*$ sequentially compact if it is weak$*$ sequentially closed. In the case $Q = Q_3$, the fact that Q is weak$*$ sequentially closed follows from Lemma 2.13 in [9], since the sets $\{w \in \mathbf{C} | \, |w| \le L^2, \Im w \ge \lambda_1\}$ and $\{w \in \mathbf{C} | \, |w| \le L^2, \Im w \ge 0, \ \Im w \ge \lambda_2 |w - k_*^2|\}$, for $k_* = k_+$ and k_-, are compact and convex.

In the case $Q = Q_4$, in order to see that Q is weak$*$ sequentially compact consider a sequence $(\mu_j) \subset Q$. Since (μ_j) is bounded, it follows from the Alaoglu theorem [16, p. 60] that there is an element $\mu \in L_\infty(\mathbf{R}^2)$ and a subsequence of (μ_j), denoted simply by itself, such that (μ_j) converges weak$*$ to μ in $L_\infty(\mathbf{R}^2)$ and $\|\mu\|_\infty \le L^2$. Thus, for all $\xi \in L_1(\mathbf{R}^2)$,

$$\int_{\mathbf{R}^2} \mu_j \xi dx \to \int_{\mathbf{R}^2} \mu \xi dx, \tag{6.8}$$

as $j \to \infty$ and, in particular, (6.8) holds if ξ is the characteristic function of any bounded measurable subset of $\mathbf{R}^2$. This and the fact that $\mu_j \in Q$, $j = 1, 2, \ldots$, implies that $\Im \mu \ge 0$ in $\mathbf{R}^2$, $\mu(x) = k_+^2$ for $x \in U_B$, $\mu(x) = k_-^2$ for $x \in \mathbf{R}^2 \backslash \overline{U}$, ess $\inf_{x \in E_\eta^\rho} \{\tilde{k}^2(x) - \Re[\mu(x)]\} \ge \lambda_2$, and

$$\begin{aligned} \text{ess inf}_{x \in U_\beta} \, \Re[\mu(x + e_2 h) - \mu(x)] &\ge 0, \\ \text{ess inf}_{x \in \mathbf{R}^2 \backslash \overline{U}_\beta} \, \Re[\mu(x - e_2 h) - \mu(x)] &\ge 0 \end{aligned}$$

for all $h > 0$. Hence $\mu \in Q$ and (μ_j) converges weak$*$ to μ in Q. Thus Q is weak$*$ sequentially compact.

Finally, let $z = (z^{(1)}, z^{(2)}) \in W$. Then, for some $\mu \in Q$, $z^{(1)} = (\mu|_{\Omega_1} - k_+^2, \ \mu|_{\Omega_2} - k_+^2, \ -1, \ 1)^t$ and $z^{(2)} = (k_+^2 - k_-^2)(1, 1, 0, 0)^t$. For $j = 1, 2, \ldots$, set

$$\mu_j(x) = \begin{cases} \mu^*(x) & \text{for } |x_1| > j, \\ \mu(x) & \text{for } |x_1| \le j, \end{cases}$$

where $\mu^* \equiv i\lambda_1$ in the case that (A3) is satisfied, $\mu^* = k_+^2$ in U_B, $= k_-^2$ in $\mathbf{R}^2 \backslash \overline{U}$, $= \min(k_-^2, k_+^2) - \lambda_2$ in E_B, in the case that (A4) and (A5) are satisfied. Then $\mu^*, \mu_j \in Q$, and setting $z_j^{(1)} = (\mu_j|_{\Omega_1} - k_+^2, \ \mu_j|_{\Omega_2} - k_+^2, \ -1, \ 1)^t$ and $z_j = (z_j^{(1)}, z^{(2)})$, $j = 1, 2, \ldots$, it is easy to see that (z_j) converges weak$*$ to z. Define

$$\begin{aligned} z^* &= ((\mu^*|_{\Omega_1} - k_+^2, \ \mu^*|_{\Omega_2} - k_+^2, \ -1, \ 1)^t, z^{(2)}) \\ &= ((\lambda^*, \lambda^*, -1, 1)^t, (k_+^2 - k_-^2, k_+^2 - k_-^2, 0, 0)^t), \end{aligned}$$

where $\lambda^* \in \mathbf{C}$ is given by $\lambda^* = i\lambda_1 - k_+^2$ in the case that (A3) is satisfied and by $\lambda^* = \min(k_-^2, k_+^2) - \lambda_2 - k_+^2$ in the case that (A4) and (A5) are satisfied. Since $z^* \in W$ so that $I - K_{z^*}$ is injective, it follows from Theorem 6.2 that $I - K_{z_j}$ injective implies $I - K_{z_j}$ surjective, for $j = 1, 2, \ldots$.

All the assumptions in Theorem 6.1 have been verified so Theorem 6.3 follows from Theorem 6.1. ∎

THEOREM 6.4. *Assume that* (A3) *holds or that* (A4) *and* (A5) *hold. Then Problem* (P) *has exactly one solution. Further, for any* $L > 0$, *there exists a constant* $C > 0$ *depending only on* L, $k_\pm$, η, ρ, B, λ_1, *and* λ_2 *in the case that* (A3) *is satisfied, or*

on L, $k_\pm$, η, ρ, B, β, *and* λ_3 *in the case that* (A4) *and* (A5) *are satisfied such that, provided* $\|k\|_\infty \le L$,

$$|u(x)| \le C(1+|x_2|)^{1/2}, \qquad x \in \mathbf{R}^2. \tag{6.9}$$

Proof. The existence of a unique solution to Problem (P) follows from Theorems 3.1, 3.2, and 6.3. To derive the estimate (6.9) we note from the equivalence of (i) and (ii) in Lemma 2.1 that, for $h > B$,

$$u^{\mathrm{s}}(x) = 2\int_{\Gamma_h} \frac{\partial \Phi(x,y)}{\partial y_2} u^{\mathrm{s}}(y)\, ds(y), \qquad x \in U_h. \tag{6.10}$$

It follows from (2.6) and (6.10) (see [5]) that

$$|u^{\mathrm{s}}(x)| \le C(1+(x_2-B))^{1/2} \sup_{x\in\Gamma_B} |u(x)|, \qquad x \in U_B, \tag{6.11}$$

for some constant $C > 0$ dependent only on k_+, which together with Theorem 6.3 implies the estimate (6.9) for $x \in U$. The estimate (6.9) for $x \in \mathbf{R}^2 \backslash U$ can be proved similarly by using Lemma 2.2. ∎

Appendix: Guided waves. By a guided wave we mean a solution of the homogeneous problem which has its energy localized in or near the layer E_B. Precisely, for $a < b$, let $D(a,b) = \{x \in \mathbf{R}^2 | a < x_1 < b\}$. Then our definition is as follows.

Definition A.1. *Call* $v \in C^1(\mathbf{R}^2)$ *a* guided wave *if* $v \not\equiv 0$, v *satisfies* (2.3), v *is bounded in* E^h_{-h} *for every* $h > 0$,

$$\sup_{n\in\mathbf{Z}} \int_{D(n,n+1)} (|v|^2 + |\nabla v|^2)\, dx < \infty, \tag{A.1}$$

and

$$c_h := \sup_{n\in\mathbf{Z}} \int_{D(n,n+1)\backslash E^h_{-h}} (|v|^2 + |\nabla v|^2)\, dx \to 0 \tag{A.2}$$

as $h \to \infty$.

Remark A.1. In the case when the scatterer is a diffraction grating, i.e., k is periodic in the x_1-direction with some period L, it is usual to assume that v is correspondingly quasi periodic (i.e., that $v(x)\exp(-ik_+ \cos\theta\, x_1)$ is periodic). Then (A.1) and (A.2) reduce to the condition that

$$\int_{D(0,L)} (|v|^2 + |\nabla v|^2)\, dx < \infty,$$

i.e., that the energy is finite in a single period of the grating (cf. Bonnet-Bendhia and Starling [3]).

Remark A.2. Conditions (A.1) and (A.2) are satisfied if v decreases rapidly enough in the vertical direction, in particular, if for some constants $C > 0$ and $p > 1/2$,

$$|v(x)| \le C(1+|x_2|)^{-p}, \qquad x \in \mathbf{R}^2.$$

The following result follows from Theorem 8.1 in [7] and Remark 2.1.

Theorem A.1. *If* v *is a guided wave, then* v *satisfies the UPRC for wavenumber* k_+ *and the DPRC for wavenumber* k_-.

REFERENCES

[1] G. Bao, *Finite element approximation of time harmonic waves in periodic structures*, SIAM J. Numer. Anal. 32 (1995), pp. 1155–1169.

[2] G. Bao, D. C. Dobson, and J. A. Cox, *Mathematical studies in rigorous grating theory*, J. Opt. Soc. Amer. A 12 (1995), pp. 1029–1042.

[3] A.-S. Bonnet-Bendhia and F. Starling, *Guided waves by electromagnetic gratings and non-uniqueness examples for the diffraction problem*, Math. Methods Appl. Sci., 17 (1994), pp. 305–338.

[4] S. N. Chandler-Wilde, *Boundary value problems for the Helmholtz equation in a half-plane*, in Proceedings, Third Int. Conf. on Mathematical and Numerical Aspects of Wave Propagation, G. Cohen, L. Halpern, and P. Joly, eds., Proceedings Appl. Math 50, SIAM, Philadelphia, PA, 1995, pp. 188–197.

[5] S. N. Chandler-Wilde, *The impedance boundary value problem for the Helmholtz equation in a half-plane*, Math. Methods Appl. Sci., 20 (1997), pp. 813–840.

[6] S. N. Chandler-Wilde and C. R. Ross, *Scattering by rough surfaces: The Dirichlet problem for the Helmholtz equation in a non-locally perturbed half-plane*, Math. Methods Appl. Sci., 19 (1996), pp. 959–976.

[7] S. N. Chandler-Wilde and B. Zhang, *Electromagnetic scattering by an inhomogeneous conducting or dielectric layer on a perfectly conducting plate*, Proc. Roy. Soc. London Ser. A, 454 (1998), pp. 519–542.

[8] S. N. Chandler-Wilde and B. Zhang, *A uniqueness result for scattering by infinite rough surfaces*, SIAM J. Appl. Math., 58 (1998), pp. 1774–1790.

[9] S. N. Chandler-Wilde and B. Zhang, *On the solvability of a class of second kind integral equations on unbounded domains*, J. Math. Anal. Appl., 214 (1997), pp. 482–502.

[10] S. N. Chandler-Wilde and B. Zhang, *A Generalized Collectively Compact Operator Theory with an Application to Integral Equations on Unbounded Domains*, in preparation.

[11] X. Chen and A. Friedman, *Maxwell's equations in a periodic structure*, Trans. Amer. Math. Soc., 323 (1991), pp. 465–507.

[12] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[13] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory,* 2nd ed., Springer-Verlag, Berlin, 1998.

[14] C. Macaskill and P. Cao, *A new treatment of rough surface scattering,* Proc. Roy. Soc. London Ser A, 452 (1996), pp. 2593–2612.

[15] J. A. DeSanto and P. A. Martin, *On the derivation of boundary integral equations for scattering by an infinite one-dimensional rough surface*, J. Acoust. Soc. Am. 102 (1997), pp. 67–77.

[16] C. L. DeVito, *Functional Analysis*, Academic Press, New York, 1978.

[17] D. Dobson and A. Friedman, *The time-harmonic Maxwell equations in a doubly periodic structure*, J. Math. Anal. Appl. 166 (1992), pp. 507–528.

[18] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[19] A. Kirsch, *Diffraction by periodic structures*, in Inverse Problems in Mathematical Physics, L. Paivarinta and E. Somersalo, eds., Lecture Notes in Phys. 422, Springer-Verlag, 1993, pp. 87–102.

[20] A. Kirsch, *An inverse problem for periodic structures,* in Inverse Scattering and Potential Problems in Mathematical Physics, R.E. Kleinman, R. Kress, and E. Martensen, eds., Peter Lang, Frankfurt, 1995, pp. 75–93.

[21] D. A. Kapp and G. S. Brown, *A new numerical method for rough surface scattering calculations*, IEEE Trans. Antennas and Propagation, 44 (1996), pp. 711–721.

[22] J. C. Nedelec and F. Starling, *Integral equation methods in a quasi-periodic diffraction problem for the time-harmonic Maxwell's equations*, SIAM J. Math. Anal. 22 (1991), pp. 1679–1701.

[23] R. Petit, *Electromagnetic Theory of Gratings*, Springer-Verlag, Berlin, 1980.

[24] H. W. Schürmann, V. S. Serov, and Yu. V. Shestopalov, *On the theory of TE-polarized waves in a linear three-layer structure*, Electromagnetic Waves and Electronic Systems, 1 (1997), pp. 49–59.

[25] B. Strycharz, *An acoustic scattering problem for a periodic, inhomogeneous media*, Math. Methods Appl. Sci., 21 (1998), pp. 969–983.

[26] L. Tsang, C. H. Chan, K. Pak, and H. Sangani, *Monte-Carlo simulations of large-scale problems of random rough surface scattering and applications to grazing incidence with*

the BMIA/canonical grid method, IEEE Trans. Antennas and Propagation, 43 (1995), pp. 851–859.

[27] G. Vainikko, *Multidimensional Weakly Singular Integral Equations*, Springer-Verlag, Berlin, 1993.

[28] G. Vainikko, *Fast Solvers of the Lippmann-Schwinger Equation*, Research reports A387, Institute of Mathematics, Helsinki University of Technology, Finland 1997.

[29] R. L. Wagner, J. M. Song, and W. C. Chew, *Monte-Carlo simulation of electromagnetic scattering from two-dimensional random rough surfaces*, IEEE Trans. Antennas and Propagation, 45 (1997), pp. 235–246.

[30] Y. Xu, *Radiation condition and scattering problem for time-harmonic acoustic waves in a stratified medium with a nonstratified inhomogeneity*, IMA J. Appl. Math., 54 (1995), pp. 9–29.

[31] Bo Zhang and S. N. Chandler-Wilde, *Acoustic scattering by an inhomogeneous layer on a rigid plate*, SIAM J. Appl. Math. 58 (1998), pp. 1931–1950.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 584–602

ON THE CLASSICAL SOLVABILITY OF THE STEFAN PROBLEM IN A VISCOUS INCOMPRESSIBLE FLUID FLOW*

YOSHIAKI KUSAKA† AND ATUSI TANI†

Abstract. This paper is devoted to the study of a solidification/melting process in the case where the fluid is flowing. Such a phenomenon is described by the Stefan problem with transport terms in the equation of the temperature distribution and the initial-boundary value problem for the incompressible Navier–Stokes equations. The existence of the classical solution is proved locally in time.

Key words. classical solvability, Stefan problem, viscous incompressible fluid

AMS subject classifications. 35R, 76D, 80

PII. S0036141098334936

1. Introduction. This paper is devoted to the study of solidification/melting processes in the case where the fluid is flowing. Such phenomena have received attention only during past three decades in physics [4], [16], [23].

In contrast to the above, we have a long history of the study of phase change in stagnant media since Stefan's pioneering work in 1891. In these problems, the unknowns are the interface separating the liquid and the solid regions and the temperature distributions in both regions. An excellent survey is provided by Rubinstein [19], Yamaguti and Nogi [25], and Meĭrmanov [15]. Meĭrmanov [14] and Hanzawa [10], in particular, studied the time-dependent, multidimensional Stefan problem in the class of smooth functions by a regularization method and by the Nash–Moser implicit function theorem, respectively.

The classical solvability of the same problem for a convective motion in a viscous incompressible fluid flow was discussed by Bazaliĭ and Degtyarev [3]. Stimulated by their work, in this paper we consider the same problem in another situation.

The differences between our problem (1.1)–(1.6) below and that of Bazaliĭ and Degtyarev consist of two aspects:

1. Equations (1.1) and (1.2) are the same except for the right-hand side of (1.2), which is caused by Boussinesq approximation in [8].
2. Bazaliĭ and Degtyarev imposed the boundary condition $\boldsymbol{v} = 0$ on the interface. Instead we impose (1.3) on the interface since, in general, the densities of a liquid and a solid at the melting temperature are not equal; for example, the densities of water and ice at 0°C are 0.999 g/cm^3 and 0.917 g/cm^3 [7] and we assume that the melting or freezing processes proceed without cavity.

Now let us formulate our problem: find the interface Γ_t which separates the liquid and the solid phases, the temperature θ, and the fluid velocity $\boldsymbol{v} = (v_1, v_2, v_3)$ satisfying

$$\frac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{v} \cdot \nabla)\boldsymbol{v} + \nabla p - \nu \Delta \boldsymbol{v} = \boldsymbol{f}(\theta), \quad \nabla \cdot \boldsymbol{v} = 0, \tag{1.1}$$

*Received by the editors March 2, 1998; accepted for publication July 21, 1998; published electronically April 7, 1999.
http://www.siam.org/journals/sima/30-3/33493.html
†Department of Mathematics, Keio University, Yokohama 223, Japan (tani@math.keio.ac.jp).

$$\frac{\partial\theta}{\partial t}+(\boldsymbol{v}\cdot\nabla)\theta-\frac{1}{\varrho C_p}\nabla\cdot(\kappa(\theta)\nabla\theta)=\frac{\nu}{2C_p}\sum_{i,k}\left(\frac{\partial v_i}{\partial x_k}+\frac{\partial v_k}{\partial x_i}\right)^2 \tag{1.2}$$

in $\bigcup_{0<t\le T}(\Omega_t\times\{t\})$, where Ω_t denotes the liquid region, ν and C_p mean kinetic viscosity and specific heat at constant pressure (assumed to be positive constants), and $\kappa(\theta)$ is heat conductivity. Equation (1.1) is known as the Navier–Stokes equations, and the right-hand side of (1.2) represents the heat generation due to viscosity.

We impose the following conditions on the interface $\bigcup_{0<t\le T}(\Gamma_t\times\{t\})$ (see, for example, [11]):

$$\boldsymbol{v}\cdot\boldsymbol{n}=\left(1-\frac{\varrho_e}{\varrho}\right)\mathcal{V},\quad 2\nu\Pi\,\mathbf{D}(\boldsymbol{v})\boldsymbol{n}=\Pi\left[\boldsymbol{v}(\boldsymbol{v}-\mathcal{V}\boldsymbol{n})^*\right]\boldsymbol{n}, \tag{1.3}$$

$$l\varrho_e\mathcal{V}=-\kappa(\theta)\nabla\theta\cdot\boldsymbol{n},\quad \theta=\theta_1. \tag{1.4}$$

Here ϱ and ϱ_e are positive constants representing the densities of liquid and solid, respectively; $\mathcal{V}$ is the normal velocity of the interface; l is a latent heat; $\mathbf{D}(\boldsymbol{v})$ is the velocity deformation tensor; Π is the projection operator to Γ_t; $\boldsymbol{n}$ is the unit normal vector to Γ_t pointing into the liquid region; θ_1 is the melting (solidification) temperature. The notation $\boldsymbol{a}^*$ is used for the transposed vector of $\boldsymbol{a}$. The conservation laws of mass and momentum imply $(1.3)^1$ and $(1.3)^2$, respectively, and (1.4) is the so-called Stefan condition which describes the conservation law of energy on the interface in the process of liquid-solid phase change.

Furthermore, in order to complete the problem, we need initial conditions and boundary conditions on the rigid boundary $\Sigma_T\equiv\Sigma\times(0,T)$:

$$\boldsymbol{v}=\boldsymbol{v}_0,\quad \theta=\theta_0\qquad\text{on}\quad \Omega\equiv\Omega_0, \tag{1.5}$$

$$\boldsymbol{v}=0,\quad \theta=\theta_2\qquad\text{on}\quad \Sigma_T. \tag{1.6}$$

Before stating our result, we introduce the function spaces used throughout this paper. Let D be a domain in $\mathbf{R}^3\times(0,T)$, l be a nonnegative integer, and $0<\alpha<1$. We denote by $C^{l+\alpha,\frac{l+\alpha}{2}}(D)$ the standard anisotropic Hölder space whose norm is $|\cdot|_D^{(l+\alpha,\frac{l+\alpha}{2})}$, and by $\tilde{C}^{3+\alpha,\frac{3+\alpha}{2}}(D)$ the function space

$$\left\{f\;\middle|\; f\in C^{3+\alpha,\frac{3+\alpha}{2}}(D),\frac{\partial f}{\partial t}\in C^{2+\alpha,\frac{2+\alpha}{2}}(D)\right\}$$

equipped with the norm

$$\|f\|_D^{(3+\alpha,\frac{3+\alpha}{2})}\equiv|f|_D^{(3+\alpha,\frac{3+\alpha}{2})}+\left|\frac{\partial f}{\partial t}\right|_D^{(2+\alpha,\frac{2+\alpha}{2})}.$$

We also define function spaces $C_0^{l+\alpha,\frac{l+\alpha}{2}}(D)$ and $\tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(D)$ as

$$\left\{f\in C^{l+\alpha,\frac{l+\alpha}{2}}(D)\;\middle|\;\left.\frac{\partial kf}{\partial tk}\right|_{t=0}=0\quad\text{for}\quad k=0,1,\dots,\left[\frac{l}{2}\right]\right\}$$

and

$$\left\{f\in\tilde{C}^{3+\alpha,\frac{3+\alpha}{2}}(D)\;\middle|\;\left.\frac{\partial kf}{\partial tk}\right|_{t=0}=0\quad\text{for}\quad k=0,1,2\right\},$$

respectively.

The following is our main result in this paper.

THEOREM 1.1. *Let T be an arbitrary positive real number and $0 < \alpha < 1$. Assume that*

$$\Gamma \equiv \Gamma_0 \in C^{3+\alpha}, \quad \Sigma \in C^{3+\alpha},$$

$$\boldsymbol{f} \in C^{1+\alpha}(0,\infty), \quad \kappa \in C^{2+\alpha}(0,\infty), \quad \boldsymbol{v}_0 \in C^{2+\alpha}(\overline{\Omega}),$$

$$\theta_0 \in C^{3+\alpha}(\overline{\Omega}), \quad \theta_1 \in C^{3+\alpha,\frac{3+\alpha}{2}}(\mathbf{R}^3 \times [0,T]), \quad \theta_2 \in C^{3+\alpha,\frac{3+\alpha}{2}}(\Sigma_T),$$

$$\kappa_0 \le \kappa(\theta) \le {\kappa_0}^{-1}, \quad \theta_2 \ge a_0, \quad |\boldsymbol{n} \cdot \nabla\theta_0|_{\Gamma_T}| \ge a_0$$

for some positive constants $\kappa_0 (< 1)$ and a_0. Moreover, we assume that the compatibility conditions up to order 1 hold. Then problem (1.1)–(1.6) has a solution

$$\Gamma_t \in \tilde{C}^{3+\alpha',\frac{3+\alpha'}{2}}, \quad \theta \in C^{3+\alpha',\frac{3+\alpha'}{2}}\left(\bigcup_{0\le t\le T_0}(\Omega_t \times \{t\})\right),$$

$$\boldsymbol{v} \in C^{2+\alpha',\frac{2+\alpha'}{2}}\left(\bigcup_{0\le t\le T_0}(\Omega_t \times \{t\})\right), \quad \nabla p \in C^{\alpha',\frac{\alpha'}{2}}\left(\bigcup_{0\le t\le T_0}(\Omega_t \times \{t\})\right)$$

for some $\alpha', 0 < \alpha' < \alpha$ and some $T_0, 0 < T_0 < T$.

2. Reduction of the problem. In this section, using the transformation due to Hanzawa [10] (see also [6]), we reduce problem (1.1)–(1.6) to that in a fixed domain. First we introduce the local coordinates $(\omega_1, \omega_2, \lambda)$ in a neighborhood of Γ, where $\omega = (\omega_1, \omega_2)$ denotes a point on the surface Γ, and a mapping x defined by $x(\omega, \lambda) = \omega + \boldsymbol{n}(\omega)\lambda$ from $\Gamma \times [-\gamma_0, \gamma_0]$ to $N_0 \subset \mathbf{R}^3$ with $\boldsymbol{n}(\omega)$ being the unit normal to Γ at ω directing into Ω. Here a positive number γ_0 is assumed to be chosen so small that the mapping x is regular and one to one. By $(\omega(x), \lambda(x))$ we denote the inverse mapping of x from N_0 to $\Gamma \times [-\gamma_0, \gamma_0]$.

Now let us assume that the interface $\Gamma_t, t \in [0,T]$ is represented by $x = \omega + \boldsymbol{n}(\omega)d(\omega,t)$ with a function $d(\omega,t)$ satisfying $d(\omega,0) = 0$. Certainly we get another representation of $\bigcup_{0\le t\le T}(\Gamma_t \times \{t\})$,

$$\bigcup_{0\le t\le T}(\Gamma_t \times \{t\}) = \left\{(x,t) \in N_0 \times [0,T] \;\middle|\; \Phi_d(x,t) = 0\right\},$$

for a function

$$\Phi_d(x,t) = \lambda(x) - d(\omega(x),t), \qquad (x,t) \in N_0 \times [0,T].$$

Accordingly, the Stefan condition can be written as

$$\frac{\partial \Phi_d}{\partial t} - c_0(\nabla\Phi_d \cdot \nabla\theta) = 0 \qquad \text{on} \qquad \bigcup_{0<t\le T}(\Gamma_t \times \{t\}),$$

where $c_0 = \kappa(\theta_1)/(l\varrho_e)$.

Next let X_T and Y_T be two coordinates (x_1, x_2, x_3, t) and (y_1, y_2, y_3, t) in $\mathbf{R}^3 \times [0, T]$ such that

$$x = (\omega(x), \lambda(x)), \qquad y = (\omega(y), \eta(y)),$$

$$\omega(x) = \omega(y), \quad \lambda(x) = \eta(y) + \chi(\eta(y))d(\omega(y), t) \qquad \text{if} \quad (x,t) \in N_0 \times [0,T],$$

$$\omega(x) = \omega(y), \quad \lambda(x) = \eta(y) \qquad \text{if} \quad (x,t) \in {N_0}^c \times [0,T],$$

where $\chi(\lambda) \in C^\infty(-\infty, +\infty)$ is a cut-off function satisfying

$$\chi(\lambda) = \begin{cases} 1 & \text{for} \quad |\lambda| \le \dfrac{\gamma_0}{4}, \\ 0 & \text{for} \quad |\lambda| \ge \dfrac{3\gamma_0}{4}, \end{cases} \qquad |\chi'(\lambda)| \le \frac{4}{3\gamma_0}.$$

Then we define a mapping $e_d : Y_T \to X_T$ by

$$e_d(y(\omega,\eta),t) = (x(\omega, \eta + \chi(\eta)d(\omega,t)), t) \qquad \text{for} \quad (x,t) \in N_0 \times [0,T],$$

$$e_d(y(\omega,\eta),t) = (x(\omega,\eta), t) \qquad \text{for} \quad (x,t) \in {N_0}^c \times [0,T].$$

It is obvious that $Q_T = \Omega \times [0,T]$ and $\Gamma_T = \Gamma \times [0,T]$ are transformed onto $\bigcup_{0\le t\le T}(\Omega_t \times \{t\})$ and $\bigcup_{0\le t\le T}(\Gamma_t \times \{t\})$, respectively, by e_d. Denoting simply by $\theta, \boldsymbol{v}$, and p the transformed functions $\theta \circ e_d, \boldsymbol{v} \circ e_d$, and $p \circ e_d$, respectively, we reduce problem (1.1)–(1.6) to that in the fixed domain Q_T in variables (y,t):

(2.1)
$$\begin{cases} \dfrac{\partial \boldsymbol{v}}{\partial t} + (\boldsymbol{h}_d \cdot \nabla)\boldsymbol{v} + (\boldsymbol{v} \cdot \nabla_d)\boldsymbol{v} - \nu {\nabla_d}^2 \boldsymbol{v} + \nabla_d p = \boldsymbol{f}(\theta), \quad \nabla_d \cdot \boldsymbol{v} = 0 \quad \text{in} \quad Q_T, \\ \boldsymbol{v}|_{t=0} = \boldsymbol{v}_0 \quad \text{on} \quad \bar{\Omega}, \\ \boldsymbol{v} \cdot \boldsymbol{n}_d = \left(1 - \dfrac{\varrho_e}{\varrho}\right)\mathcal{V}, \quad 2\nu \Pi_d \mathbf{D}_d(\boldsymbol{v})\boldsymbol{n}_d = \Pi_d[\boldsymbol{v}(\boldsymbol{v} - \mathcal{V}\boldsymbol{n}_d)^*]\boldsymbol{n}_d \quad \text{on} \quad \Gamma_T, \\ \boldsymbol{v} = 0 \quad \text{on} \quad \Sigma_T, \end{cases}$$

(2.2)
$$\begin{cases} \dfrac{\partial \theta}{\partial t} + (\boldsymbol{h}_d \cdot \nabla)\theta + (\boldsymbol{v} \cdot \nabla_d)\theta - \dfrac{1}{\varrho C_p}\nabla_d \cdot (\kappa(\theta)\nabla_d \theta) \\ \qquad = \dfrac{\nu}{2C_p} \displaystyle\sum_{i,k} \left(\sum_j \left\{ a^{kj} \frac{\partial v_i}{\partial y_j} + a^{ij} \frac{\partial v_k}{\partial y_j} \right\} \right)^2 \qquad \text{in} \quad Q_T, \\ \theta|_{t=0} = \theta_0 \qquad \text{on} \quad \bar{\Omega}, \\ \theta = \theta_1, \quad \dfrac{\partial d}{\partial t} + c_0(\nabla_d \Phi_d \cdot \nabla_d \theta) = 0 \quad \text{on} \quad \Gamma_T, \\ d(\omega, 0) = 0 \quad \text{on} \quad \Gamma, \\ \theta = \theta_2 \quad \text{on} \quad \Sigma_T. \end{cases}$$

Here we use the notation

$$\nabla_d = (E_d^*)^{-1}\nabla, \quad \boldsymbol{h}_d = \frac{\partial y}{\partial t} \circ e_d, \quad \boldsymbol{n}_d = \frac{\nabla_d \eta}{|\nabla_d \eta|}, \quad \mathbf{D}_d(\boldsymbol{v}) = \mathbf{D}(\boldsymbol{v}) \circ e_d, \quad \Pi_d g = \Pi g \circ e_d;$$

$E_d = (a_{ij})$ is the Jacobian matrix of the mapping from y to x, E_d^* is the transposed matrix of E_d, and a^{ij} is the (i,j)-component of $(E_d^*)^{-1}$.

It is obvious that there exist the extensions $\hat{\theta} \in C^{3+\alpha,\frac{3+\alpha}{2}}(Q_T)$, $\hat{d} \in C^{4+\alpha,\frac{4+\alpha}{2}}(\Gamma_T)$, $\hat{\boldsymbol{v}} \in C^{2+\alpha,\frac{2+\alpha}{2}}(Q_T)$, $\nabla\hat{p} \in C^{\alpha,\frac{\alpha}{2}}(Q_T)$ which satisfy the conditions

$$
\begin{cases}
\hat{\theta}(y,0) = \theta_0(y), \quad \dfrac{\partial\hat{\theta}}{\partial t}(y,0) = \theta^{(1)}(y), \\
\hat{d}(\omega,0) = 0, \quad \dfrac{\partial\hat{d}}{\partial t}(\omega,0) = d^{(1)}(\omega), \quad \dfrac{\partial^2\hat{d}}{\partial t^2}(\omega,0) = d^{(2)}(\omega), \\
\hat{\boldsymbol{v}}(y,0) = \boldsymbol{v}_0(y), \quad \dfrac{\partial\hat{\boldsymbol{v}}}{\partial t}(y,0) = \boldsymbol{v}^{(1)}(y), \\
\nabla\hat{p}(y,0) = -\boldsymbol{v}^{(1)}(y) - (\boldsymbol{v}_0(y)\cdot\nabla)\boldsymbol{v}_0(y) + \nu\nabla^2\boldsymbol{v}_0(y) + \boldsymbol{f}(\theta_0),
\end{cases}
$$

where $\theta^{(1)}, d^{(1)}, d^{(2)}$, and $\boldsymbol{v}^{(1)}$ are defined from the compatibility conditions of the equations and data in (2.1) and (2.2). It is clear that such extensions can be taken to satisfy the inequality

$$
|\hat{\theta}|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})} + |\hat{d}|_{\Gamma_T}^{(4+\alpha,\frac{4+\alpha}{2})} + |\nabla\hat{p}|_{Q_T}^{(\alpha,\frac{\alpha}{2})} + |\hat{\boldsymbol{v}}|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})} \le C\left(|\theta_0|_{\bar{\Omega}}^{(3+\alpha)} + |\boldsymbol{v}_0|_{\bar{\Omega}}^{(2+\alpha)}\right)
$$

with a constant C being bounded as $T \to 0$.

Then by introducing the new functions $w \equiv \theta - \hat{\theta} - \chi\sigma\partial\hat{\theta}/\partial n, \sigma \equiv d - \hat{d}, \boldsymbol{u} \equiv \boldsymbol{v} - \hat{\boldsymbol{v}}$ and $\nabla q \equiv \nabla p - \nabla\hat{p}$, problems (2.1) and (2.2) can be written in the equivalent form

$$
\begin{cases}
\dfrac{\partial\boldsymbol{u}}{\partial t} - \nu\Delta\boldsymbol{u} + \nabla q = \boldsymbol{f}\left(w + \hat{\theta} + \chi\dfrac{\partial\hat{\theta}}{\partial n}\sigma\right) - \dfrac{\partial\hat{\boldsymbol{v}}}{\partial t} - \left(\boldsymbol{h}_{\sigma+\hat{d}}\cdot\nabla\right)(\boldsymbol{u}+\hat{\boldsymbol{v}}) \\
\qquad +\nu\left(\nabla_{\sigma+\hat{d}}^2 - \nabla^2\right)(\boldsymbol{u}+\hat{\boldsymbol{v}}) + \nu\Delta\hat{\boldsymbol{v}} - \left((\boldsymbol{u}+\hat{\boldsymbol{v}})\cdot\nabla_{\sigma+\hat{d}}\right)(\boldsymbol{u}+\hat{\boldsymbol{v}}) \\
\qquad -\nabla_{\sigma+\hat{d}}(q+\hat{p}) + \nabla q \equiv \mathcal{F}_1(\boldsymbol{u},\nabla q, w, \sigma) \quad \text{in} \quad Q_T, \\
\nabla\cdot\boldsymbol{u} = -\left(\nabla_{\sigma+\hat{d}} - \nabla\right)\cdot\boldsymbol{u} - \nabla_{\sigma+\hat{d}}\cdot\hat{\boldsymbol{v}} \equiv \mathcal{F}_2(\boldsymbol{u},\sigma) \quad \text{in} \quad Q_T, \\
\boldsymbol{u}\,\big|_{t=0} = 0 \quad \text{on} \quad \bar{\Omega}, \\
\boldsymbol{u}\cdot\boldsymbol{n} = -\boldsymbol{u}\cdot\left(\boldsymbol{n}_{\sigma+\hat{d}} - \boldsymbol{n}\right) - \hat{\boldsymbol{v}}\cdot\boldsymbol{n}_{\sigma+\hat{d}} + \left(1 - \dfrac{\varrho_e}{\varrho}\right)\left(\dfrac{\partial\sigma}{\partial t} + \dfrac{\partial\hat{d}}{\partial t}\right) \\
\qquad\qquad \equiv \mathcal{F}_3(\boldsymbol{u},\sigma) \quad \text{on} \quad \Gamma_T, \\
2\nu\Pi\mathbf{D}(\boldsymbol{u})\boldsymbol{n} = -\left(2\nu\Pi_{\sigma+\hat{d}}\mathbf{D}_{\sigma+\hat{d}}(\boldsymbol{u}+\hat{\boldsymbol{v}})\boldsymbol{n}_{\sigma+\hat{d}} - 2\nu\Pi\mathbf{D}(\boldsymbol{u}+\hat{\boldsymbol{v}})\boldsymbol{n}\right) \\
\qquad -2\nu\Pi\mathbf{D}(\hat{\boldsymbol{v}})\boldsymbol{n} + \Pi_{\sigma+\hat{d}}\left[(\boldsymbol{u}+\hat{\boldsymbol{v}})\left(\boldsymbol{u}+\hat{\boldsymbol{v}} - \dfrac{\partial}{\partial t}(\sigma+\hat{d})\boldsymbol{n}_{\sigma+\hat{d}}\right)^*\right]\boldsymbol{n}_{\sigma+\hat{d}} \\
\qquad\qquad \equiv \mathcal{F}_4(\boldsymbol{u},\sigma) \quad \text{on} \quad \Gamma_T, \\
\boldsymbol{u} = -\hat{\boldsymbol{v}} \quad \text{on} \quad \Sigma_T,
\end{cases}
\tag{2.3}
$$

$$
\begin{cases}
\dfrac{\partial w}{\partial t} - \dfrac{1}{\varrho C_p}\nabla\cdot(\kappa(\hat\theta)\nabla w) = -\dfrac{\partial}{\partial t}\left(\hat\theta + \chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right) \\
\quad -(\boldsymbol{h}_{\sigma+\hat d}\cdot\nabla)\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right) - \left((\boldsymbol{u}+\hat{\boldsymbol{v}})\cdot\nabla_{\sigma+\hat d}\right)\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right) \\
\quad +\dfrac{1}{\varrho C_p}\nabla\cdot\left(\kappa(\hat\theta)\nabla\left(\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\right) \\
\quad -\dfrac{1}{\varrho C_p}\nabla\cdot\left(\kappa(\hat\theta)\nabla\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\right) \\
\quad +\dfrac{1}{\varrho C_p}\nabla_{\sigma+\hat d}\cdot\left(\kappa\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\nabla_{\sigma+\hat d}\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\right) \\
\quad +\dfrac{\nu}{2C_p}\sum_{i,k}\left(\sum_j\left\{a^{kj}\dfrac{\partial(u_i+\hat v_i)}{\partial y_j}+a^{ij}\dfrac{\partial(u_k+\hat v_k)}{\partial y_j}\right\}\right)^2 \\
\qquad \equiv \mathcal{F}_5(\boldsymbol{u},w,\sigma) \quad \text{in} \quad Q_T, \\
w\,\big|_{t=0} = 0 \quad \text{on} \quad \bar\Omega, \\
w+\dfrac{\partial\hat\theta}{\partial n}\sigma = \theta_1-\hat\theta \quad \text{on} \quad \Gamma_T, \\
\dfrac{\partial\sigma}{\partial t}+c_0(\nabla\eta\cdot\nabla w) = -\dfrac{\partial\hat d}{\partial t} - c_0\left(\nabla_{\sigma+\hat d}\eta\cdot\nabla_{\sigma+\hat d}\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\right) \\
\quad +c_0\left(\nabla\eta\cdot\nabla\left(w+\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\right) - c_0\left(\nabla\eta\cdot\nabla\left(\hat\theta+\chi\dfrac{\partial\hat\theta}{\partial n}\sigma\right)\right) \\
\qquad \equiv \mathcal{F}_6(w,\sigma) \quad \text{on} \quad \Gamma_T, \\
\sigma\,\big|_{t=0} = 0 \quad \text{on} \quad \Gamma, \\
w = \theta_2-\hat\theta \quad \text{on} \quad \Sigma_T.
\end{cases} \tag{2.4}
$$

3. Linear problems. In this section we consider the linear problems

$$
\begin{cases}
\dfrac{\partial\boldsymbol{u}}{\partial t} - \nu\Delta\boldsymbol{u} + \nabla p = \boldsymbol{F}_1, \quad \nabla\cdot\boldsymbol{u} = F_2 \qquad \text{in} \quad Q_T, \\
\boldsymbol{u}\,\big|_{t=0} = 0 \qquad \text{on} \quad \bar\Omega, \\
\boldsymbol{u}\cdot\boldsymbol{n} = F_3, \quad \Pi\mathbf{D}(\boldsymbol{u})\boldsymbol{n} = \boldsymbol{F}_4 \qquad \text{on} \quad \Gamma_T, \\
\boldsymbol{u} = \boldsymbol{G}_1 \qquad \text{on} \quad \Sigma_T,
\end{cases} \tag{3.1}
$$

$$
\begin{cases}
\dfrac{\partial w}{\partial t} - \dfrac{1}{\varrho C_p}\nabla\cdot(\kappa(\hat\theta)\nabla w) = F_5 \qquad \text{in} \quad Q_T, \\
w\,\big|_{t=0} = 0 \qquad \text{on} \quad \bar\Omega, \\
w+\dfrac{\partial\hat\theta}{\partial n}\sigma = F_6, \quad \dfrac{\partial\sigma}{\partial t}+c_0\dfrac{\partial w}{\partial n} = F_7 \qquad \text{on} \quad \Gamma_T, \\
w = G_2 \qquad \text{on} \quad \Sigma_T.
\end{cases} \tag{3.2}
$$

We will solve problems (3.1) and (3.2) separately. First we consider the following model problem in the half-space $D^3_\infty = R^3_+ \times (0,\infty)$, $R^3_+ = \{(z_1,z_2,z_3) \in \mathbf{R}^3 \mid z_3 > 0\}$.

$$\text{(3.3)} \qquad \begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} - \nu\Delta \boldsymbol{u} + \nabla p = 0, \quad \nabla \cdot \boldsymbol{u} = 0 & \text{in} \quad D^3_\infty, \\ \boldsymbol{u}\,\big|_{t=0} = 0 \quad \text{on} \quad R^3_+, \\ \dfrac{\partial u_3}{\partial z_j} + \dfrac{\partial u_j}{\partial z_3} = b_j \quad (j=1,2), \quad u_3 = 0 & \text{on} \quad R^2_\infty, \end{cases}$$

where $b_j, j = 1,2$, are given functions defined on $R^2_\infty = \mathbf{R}^2 \times (0,\infty)$.

By means of Fourier transformation with respect to $z' = (z_1, z_2)$ and Laplace transformation with respect to t,

$$[\mathcal{F}f](\xi', z_3, s) \equiv \tilde{f}(\xi', z_3, s) = \int_0^\infty \mathrm{e}^{-st}dt \int_{\mathbf{R}^2} \mathrm{e}^{-iz'\cdot\xi'} f(z', z_3, t)dz',$$

the solution of the transformed problem of (3.3) can be represented as

$$\text{(3.4)} \qquad \begin{cases} \tilde{u}_j = \dfrac{e_0(z_3)}{r} + \dfrac{i\xi_j}{|\xi'|(r+|\xi'|)} \displaystyle\sum_{k=1,2} i\xi_k \tilde{b}_k \left(e_0(z_3) - e_1(z_3)\right), \quad j = 1,2, \\ \tilde{u}_3 = \dfrac{e_1(z_3)}{r+|\xi'|} \displaystyle\sum_{k=1,2} i\xi_k \tilde{b}_k, \quad \tilde{p} = -\dfrac{\nu}{|\xi'|} e_2(z_3) \displaystyle\sum_{k=1,2} i\xi_k \tilde{b}_k, \end{cases}$$

where

$$e_0(z_3) = e^{-rz_3}, \quad e_1(z_3) = \frac{e^{-rz_3} - e^{-|\xi'|z_3}}{r - |\xi'|}, \quad e_2(z_3) = e^{-|\xi'|z_3},$$

$$r^2 = \frac{s}{\nu} + \xi'^2, \quad s = a + i\xi_0, \quad \xi'^2 = \xi_1^2 + \xi_2^2 \quad \text{(see [24])}.$$

From (3.4) we can easily obtain

$$\text{(3.5)} \qquad |\boldsymbol{u}|^{(2+\alpha,\frac{2+\alpha}{2})}_{D^3_T} + |\nabla p|^{(\alpha,\frac{\alpha}{2})}_{D^3_T} \le C \sum_{k=1,2} |b_k|^{(1+\alpha,\frac{1+\alpha}{2})}_{R^2_T},$$

where $D^3_T = R^3_+ \times (0,T)$, $R^2_T = \mathbf{R}^2 \times (0,T)$.

Second we proceed to the nonhomogeneous problem

$$\text{(3.6)} \qquad \begin{cases} \dfrac{\partial \boldsymbol{u}}{\partial t} - \nu\Delta \boldsymbol{u} + \nabla p = \boldsymbol{f}, \quad \nabla \cdot \boldsymbol{u} = g & \text{in} \quad D^3_T, \\ \boldsymbol{u}\,\big|_{t=0} = 0 \quad \text{on} \quad R^3_+, \\ \dfrac{\partial u_3}{\partial z_j} + \dfrac{\partial u_j}{\partial z_3}\,\Big|_{z_3=0} = b_j \quad (j=1,2), \quad u_3\,\big|_{z_3=0} = b_3 & \text{on} \quad R^2_T. \end{cases}$$

It is obvious that the solution of problem (3.6) is given by

$$\boldsymbol{u} = \boldsymbol{u}' + \nabla\phi + \boldsymbol{u}'', \quad p = \pi - \frac{\partial \phi}{\partial t} + \nu\Delta\phi,$$

where $\boldsymbol{u}'$ is a solution to the Dirichlet problem of the heat equation

$$\begin{cases} \dfrac{\partial \boldsymbol{u}'}{\partial t} - \nu \Delta \boldsymbol{u}' = \boldsymbol{f} & \text{in} \quad D_T^3, \\ \boldsymbol{u}' \,\big|_{t=0} = 0, \quad \boldsymbol{u}' \,\big|_{z_3=0} = 0 & \text{on} \quad R_T^2; \end{cases} \tag{3.7}$$

ϕ is a solution of the Neumann problem of the elliptic equation with a parameter t

$$\begin{cases} \Delta \phi = g - \nabla \cdot \boldsymbol{u}' & \text{in} \quad R_+^3, \\ \dfrac{\partial \phi}{\partial z_3} \,\big|_{z_3=0} = b_3 & \text{on} \quad \mathbf{R}^2; \end{cases} \tag{3.8}$$

and $(\boldsymbol{u}'', \pi)$ is a solution of problem

$$\begin{cases} \dfrac{\partial \boldsymbol{u}''}{\partial t} - \nu \Delta \boldsymbol{u}'' + \nabla \pi = 0, \quad \nabla \cdot \boldsymbol{u}'' = 0 \quad \text{in} \quad D_T^3, \\ \boldsymbol{u}'' \,\big|_{t=0} = 0 \quad \text{on} \quad R_+^3, \\ \left(\dfrac{\partial u_3''}{\partial z_j} + \dfrac{\partial u_j''}{\partial z_3} \right) \big|_{z_3=0} = b_j - \dfrac{\partial u_3'}{\partial z_j} - \dfrac{\partial u_j'}{\partial z_3} - 2 \dfrac{\partial^2 \phi}{\partial z_3 \partial z_j} \,\big|_{z_3=0} \quad (j = 1, 2), \\ u_3'' \,\big|_{z_3=0} = 0 \quad \text{on} \quad R_T^2. \end{cases} \tag{3.9}$$

It is well known that problems (3.7) and (3.8) have solutions satisfying

$$|\boldsymbol{u}'|_{D_T^3}^{(2+\alpha, \frac{2+\alpha}{2})} \le C \, |\boldsymbol{f}|_{D_T^3}^{(\alpha, \frac{\alpha}{2})},$$

$$|\nabla \phi|_{R_+^3}^{(2+\alpha)} \le C \left(|g|_{R_+^3}^{(1+\alpha)} + |\boldsymbol{u}'|_{R_+^3}^{(2+\alpha)} + |b_3|_{\mathbf{R}^2}^{(2+\alpha)} \right).$$

With regard to problem (3.9), it follows from (3.5) that

$$\begin{aligned} &|\boldsymbol{u}''|_{D_T^3}^{(2+\alpha, \frac{2+\alpha}{2})} + |\nabla \pi|_{D_T^3}^{(\alpha, \frac{\alpha}{2})} \\ &\le C \left(\sum_{j=1,2} |b_j|_{R_T^2}^{(1+\alpha, \frac{1+\alpha}{2})} + |\boldsymbol{u}'|_{R_T^2}^{(2+\alpha, \frac{2+\alpha}{2})} + |\nabla \phi|_{R_T^2}^{(2+\alpha, \frac{2+\alpha}{2})} \right). \end{aligned}$$

Thus the solution of problem (3.6) is evaluated as follows:

$$|\boldsymbol{u}|_{D_T^3}^{(2+\alpha, \frac{2+\alpha}{2})} + |\nabla p|_{D_T^3}^{(\alpha, \frac{\alpha}{2})} \le C \left(|\boldsymbol{f}|_{D_T^3}^{(\alpha, \frac{\alpha}{2})} + |g|_{D_T^3}^{(1+\alpha, \frac{1+\alpha}{2})} + \sum_{j=1,2} |b_j|_{R_T^2}^{(1+\alpha, \frac{1+\alpha}{2})} \right).$$

Finally we solve problem (3.1) by a regularizer. Let $\{\omega^{(k)}\}$ and $\{\Omega^{(k)}\}$ be two systems of the coverings of $\bar{\Omega}$ constructed in the same way as in [12] and let λ be a positive constant determined later. For $k \in M_1$, $\omega^{(k)}$ and $\Omega^{(k)}$ are three-dimensional cubes included completely in Ω with common centers and with the length of their edges, in parallel directions of axes, equal to $\lambda/2$ and λ, respectively. For $k \in M_2, \omega^{(k)}$ and $\Omega^{(k)}$ have common parts with Σ which are defined in the local coordinates $\{z\}$ in the neighborhood of Σ as follows:

$$\omega^{(k)} = \Pi_x^z \left\{ |z_i| \le \frac{\lambda}{2} \quad (i = 1, 2), \quad 0 \le z_3 - F(z_1, z_2) \le \lambda \right\}$$

and

$$\Omega^{(k)} = \Pi_x^z \left\{ |z_i| \le \lambda \quad (i = 1, 2), \quad 0 \le z_3 - F(z_1, z_2) \le 2\lambda \right\},$$

respectively, where $z_3 = F(z_1, z_2)$ represents Σ and Π_x^z is the transformation from z to x. For $k \in M_3$, $\omega^{(k)}$ and $\Omega^{(k)}$, which are adjacent to Γ, are defined in the same way as $\omega^{(k)}$ and $\Omega^{(k)}$ for $k \in M_2$. Furthermore, we introduce the partitions of unity $\{\xi^{(k)}\}$ and $\{\eta^{(k)}\}$ subordinated to $\{\Omega^{(k)}\}$ and $\{\omega^{(k)}\}$ such that

$$\xi^{(k)}(x) = \begin{cases} 1 & \left(x \in \omega^{(k)}\right), \\ 0 & \left(x \in \Omega \backslash \Omega^{(k)}\right), \end{cases} \qquad 0 \le \xi^{(k)}(x) \le 1,$$

$$|D_x^\alpha \xi^{(k)}(x)| \le C\lambda^{-|\alpha|}, \quad \eta^{(k)}(x) = \frac{\xi^{(k)}(x)}{\sum_k \xi^{(k)}(x)^2}.$$

Then a regularizer $\mathcal{R}$ is defined by

$$\mathcal{R}H = \sum_{k \in M_1} \eta^{(k)}(\boldsymbol{u}_1^{(k)}, p_1^{(k)}) + \sum_{j=2,3} \sum_{k \in M_j} \eta^{(k)} \Pi_x^z (\boldsymbol{u}_j^{(k)}, p_j^{(k)}),$$

where $H = (\boldsymbol{F}_1, F_2, F_3, \boldsymbol{F}_4, \boldsymbol{G}_1)$. Here $(\boldsymbol{u}_1^{(k)}, p_1^{(k)})$ is a solution of the problem

$$\begin{cases} \dfrac{\partial \boldsymbol{u}_1^{(k)}}{\partial t} - \nu \Delta \boldsymbol{u}_1^{(k)} + \nabla p_1^{(k)} = \xi^{(k)} \boldsymbol{F}_1, \quad \nabla \cdot \boldsymbol{u}_1^{(k)} = \xi^{(k)} F_2 & \text{in} \quad R_T^3, \\ \boldsymbol{u}_1^{(k)} \big|_{t=0} = 0 & \text{on} \quad \mathbf{R}^3, \end{cases}$$

$(\boldsymbol{u}_2^{(k)}, p_2^{(k)})$ is a solution of the problem

$$\begin{cases} \dfrac{\partial \boldsymbol{u}_2^{(k)}}{\partial t} - \nu \Delta \boldsymbol{u}_2^{(k)} + \nabla p_2^{(k)} = \Pi_z^x \xi^{(k)} \boldsymbol{F}_1, \quad \nabla \cdot \boldsymbol{u}_2^{(k)} = \Pi_z^x \xi^{(k)} F_2 \quad \text{in} \quad D_T^3, \\ \boldsymbol{u}_2^{(k)} \big|_{t=0} = 0 \quad \text{on} \quad R_+^3, \\ \boldsymbol{u}_2^{(k)} \big|_{z_3=0} = \Pi_z^x \xi^{(k)} \boldsymbol{G}_1 \quad \text{on} \quad R_T^2, \end{cases}$$

and $(\boldsymbol{u}_3^{(k)}, p_3^{(k)})$ is a solution of the problem

$$\begin{cases} \dfrac{\partial \boldsymbol{u}_3^{(k)}}{\partial t} - \nu \Delta \boldsymbol{u}_3^{(k)} + \nabla p_3^{(k)} = \Pi_z^x \xi^{(k)} \boldsymbol{F}_1, \quad \nabla \cdot \boldsymbol{u}_3^{(k)} = \Pi_z^x \xi^{(k)} F_2 \quad \text{in} \quad D_T^3, \\ \boldsymbol{u}_3^{(k)} \big|_{t=0} = 0 \quad \text{on} \quad R_+^3, \\ u_{3,3}^{(k)} \big|_{z_3=0} = \Pi_z^x \xi^{(k)} F_3, \quad \dfrac{\partial u_{3,3}^{(k)}}{\partial z_j} + \dfrac{\partial u_{3,j}^{(k)}}{\partial z_3} \big|_{z_3=0} = \Pi_z^x \xi^{(k)} F_{4,j} \ \ (j = 1, 2) \quad \text{on} \quad R_T^2. \end{cases}$$

It is not difficult to see that $\mathcal{R}H = (\bar{\boldsymbol{v}}, \bar{q}), H = (\boldsymbol{F}_1, F_2, F_3, \boldsymbol{F}_4, \boldsymbol{G}_1)$ satisfies

$$\begin{cases} \dfrac{\partial \bar{\boldsymbol{v}}}{\partial t} - \nu \Delta \bar{\boldsymbol{v}} + \nabla \bar{q} = \boldsymbol{F}_1 - \mathcal{T}_1 H, \quad \nabla \cdot \bar{\boldsymbol{v}} = F_2 - \mathcal{T}_2 H \quad \text{in} \quad Q_T, \\ \bar{\boldsymbol{v}} \big|_{t=0} = 0 \quad \text{on} \quad \bar{\Omega}, \\ \bar{\boldsymbol{v}} \cdot \boldsymbol{n} = F_3 - \mathcal{T}_3 H, \quad \Pi \mathbf{D}(\bar{\boldsymbol{v}}) \boldsymbol{n} = \boldsymbol{F}_4 - \mathcal{T}_4 H \quad \text{on} \quad \Gamma_T, \\ \bar{\boldsymbol{v}} = \boldsymbol{G}_1 - \mathcal{S}_1 H \quad \text{on} \quad \Sigma_T, \end{cases}$$

where

$$
\begin{cases}
\mathcal{T}_1 H = -\nu \displaystyle\sum_{k\in M_1} \left\{\eta^{(k)}\Delta \boldsymbol{u}_1^{(k)} - \Delta(\eta^{(k)}\boldsymbol{u}_1^{(k)})\right\} \\
\qquad -\nu \displaystyle\sum_{j=2,3}\sum_{k\in M_j} \left\{\eta^{(k)}\Pi_x^z(\Delta - \overline{\nabla}^2)\boldsymbol{u}_j^{(k)} + \eta^{(k)}\Pi_x^z(\overline{\nabla}^2\boldsymbol{u}_j^{(k)}) - \Delta(\eta^{(k)}\Pi_x^z\boldsymbol{u}_j^{(k)})\right\} \\
\qquad + \displaystyle\sum_{k\in M_1} \left\{\eta^{(k)}\nabla p_1^{(k)} - \nabla(\eta^{(k)}p_1^{(k)})\right\} \\
\qquad + \displaystyle\sum_{j=2,3}\sum_{k\in M_j} \left\{\eta^{(k)}\Pi_x^z(\nabla - \overline{\nabla})p_j^{(k)} + \eta^{(k)}\Pi_x^z\overline{\nabla}p_j^{(k)} - \nabla(\eta^{(k)}\Pi_x^z p_j^{(k)})\right\}, \\
\mathcal{T}_2 H = \displaystyle\sum_{k\in M_1} \left\{\eta^{(k)}\nabla\cdot\boldsymbol{u}_1^{(k)} - \nabla\cdot(\eta^{(k)}\boldsymbol{u}_1^{(k)})\right\} \\
\qquad + \displaystyle\sum_{j=2,3}\sum_{k\in M_j} \left\{\eta^{(k)}\Pi_x^z(\nabla - \overline{\nabla})\cdot\boldsymbol{u}_j^{(k)} + \eta^{(k)}\Pi_x^z(\overline{\nabla}\cdot\boldsymbol{u}_j^{(k)}) - \nabla\cdot(\eta^{(k)}\Pi_x^z\boldsymbol{u}_j^{(k)})\right\}, \\
\mathcal{T}_3 H = \displaystyle\sum_{k\in M_3} \left\{\eta^{(k)}\Pi_x^z(\boldsymbol{u}_3^{(k)}\cdot\boldsymbol{n}_0) - (\eta^{(k)}\Pi_x^z\boldsymbol{u}_3^{(k)})\cdot\boldsymbol{n}\right\}, \\
\mathcal{T}_4 H = \displaystyle\sum_{k\in M_3} \left\{\eta^{(k)}\Pi_x^z\left(\Pi_0(\mathbf{D}(\boldsymbol{u}_3^{(k)}) - \overline{\mathbf{D}}(\boldsymbol{u}_3^{(k)}))\boldsymbol{n}_0\right)\right\} \\
\qquad + \displaystyle\sum_{k\in M_3} \left\{\eta^{(k)}\Pi_x^z\left(\Pi_0\overline{\mathbf{D}}(\boldsymbol{u}_3^{(k)})\boldsymbol{n}_0\right) - \Pi\mathbf{D}\left(\eta^{(k)}\Pi_x^z\boldsymbol{u}_3^{(k)}\right)\boldsymbol{n}\right\}, \\
\mathcal{S}_1 H = 0,
\end{cases}
$$

$\overline{\nabla} = {}^t(\partial x_i/\partial z_j)^{-1}\nabla$, $\overline{\mathbf{D}} = \Pi_z^x\mathbf{D}, \boldsymbol{n}_0 = {}^t(0,0,1)$, $\Pi_0\boldsymbol{g} = \boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{n}_0)\boldsymbol{n}_0$, and $\Pi\boldsymbol{g} = \boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{n})\boldsymbol{n}$. This means that $\mathcal{R}$ is an operator defined on the space

$$
\begin{aligned}
\mathcal{H} = {} & C_0^{\alpha,\frac{\alpha}{2}}(Q_T) \times C_0^{1+\alpha,\frac{1+\alpha}{2}}(Q_T) \\
& \times C_0^{2+\alpha,\frac{2+\alpha}{2}}(\Gamma_T) \times C_0^{1+\alpha,\frac{1+\alpha}{2}}(\Gamma_T) \times C_0^{2+\alpha,\frac{2+\alpha}{2}}(\Sigma_T)
\end{aligned}
$$

with the range $C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T) \times C_0^{\alpha,\frac{\alpha}{2}}(Q_T)$.

We claim that the norm $\|\mathcal{T}\|$ of the operator $\mathcal{T} = (\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, \mathcal{T}_4, \mathcal{S}_1)$ on $\mathcal{H}$ is less than 1. Then we can find a solution of (3.1) in the form $\mathcal{R}(I-\mathcal{T})^{-1}H$. The proof of the contraction of $\mathcal{T}$ is not difficult. Indeed, for example, evaluating each term in $\mathcal{T}_1$, we have

$$
\begin{aligned}
\sup_k |\mathcal{T}_1 H|_{Q_T^{(k)}}^{(\alpha,\frac{\alpha}{2})} &\le C\left(\lambda + \frac{T^{\frac{1}{2}}}{\lambda} + \frac{T}{\lambda^2} + T^{\frac{\alpha}{2}}\right) \sup_k \left\{|\boldsymbol{u}^{(k)}|_{Q_T^{(k)}}^{(2+\alpha,\frac{2+\alpha}{2})} + |\nabla p^{(k)}|_{Q_T^{(k)}}^{(\alpha,\frac{\alpha}{2})}\right\} \\
&\le \frac{1}{2}\|H\|_{\mathcal{H}}, \quad Q_T^{(k)} = \Omega^{(k)} \times (0,T),
\end{aligned}
$$

for sufficiently small λ and T. In the same way similar estimates for $\mathcal{T}_j H, j = 2,3,4$ and $\mathcal{S}_1$ are derived. Hence we have the following theorem.

THEOREM 3.1. *Let us assume that*

$$
\boldsymbol{F}_1 \in C_0^{\alpha,\frac{\alpha}{2}}(Q_T), F_2 \in C_0^{1+\alpha,\frac{1+\alpha}{2}}(Q_T),
$$

$$
F_3 \in C_0^{2+\alpha,\frac{2+\alpha}{2}}(\Gamma_T), \boldsymbol{F}_4 \in C_0^{1+\alpha,\frac{1+\alpha}{2}}(\Gamma_T), \boldsymbol{G}_1 \in C_0^{2+\alpha,\frac{2+\alpha}{2}}(\Sigma_T).
$$

Then problem (3.1) *has a unique solution* $\boldsymbol{u} \in C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T), \nabla q \in C_0^{\alpha,\frac{\alpha}{2}}(Q_T)$ *satisfying*

$$|\boldsymbol{u}|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})} + |\nabla q|_{Q_T}^{(\alpha,\frac{\alpha}{2})} \le C\left(|\boldsymbol{F}_1|_{Q_T}^{(\alpha,\frac{\alpha}{2})} + |F_2|_{Q_T}^{(1+\alpha,\frac{1+\alpha}{2})} + |F_3|_{\Gamma_T}^{(2+\alpha,\frac{2+\alpha}{2})} + |\boldsymbol{F}_4|_{\Gamma_T}^{(1+\alpha,\frac{1+\alpha}{2})} + |\boldsymbol{G}_1|_{\Sigma_T}^{(2+\alpha,\frac{2+\alpha}{2})}\right), \tag{3.10}$$

where a constant C *depends only on* $\boldsymbol{F}_1, F_2, F_3, \boldsymbol{F}_4, \boldsymbol{G}_1$, *and* T, *and remains bounded as* $T \to 0$.

Next we turn to problem (3.2). In the same way as above, first we treat the model problem

$$\begin{cases} \dfrac{\partial w'}{\partial t} - a\Delta w' = 0 \quad \text{in} \quad D_\infty^3, \\ w' \,\big|_{t=0} = 0 \quad \text{on} \quad R_+^3, \\ w' + b\sigma = 0 \quad \dfrac{\partial \sigma}{\partial t} + c\dfrac{\partial w'}{\partial z_3} = f \quad \text{on} \quad R_\infty^2, \end{cases} \tag{3.11}$$

where a, b, and c are positive constants, and f is a given function in $C_0^{2+\alpha,\frac{2+\alpha}{2}}(R_\infty^2)$ (see also [3]).

Making use of the Fourier–Laplace transformation, we obtain the explicit representation of the transformed unknowns

$$\tilde{w'} = -b\tilde{\sigma}\exp\left[-\left(\frac{s + a\xi'^2}{a}\right)^{1/2} z_3\right], \quad \tilde{\sigma} = \frac{\tilde{f}}{s + cb(a^{-1}s + \xi'^2)^{1/2}}. \tag{3.12}$$

Since the inverse Fourier–Laplace transformation of the symbol $(s+cb(a^{-1}s+\xi'^2)^{1/2})^{-1}$ is

$$\begin{aligned} K(z',t) &= \frac{a^{1/2}}{bc}\mathcal{F}^{-1}\left(\frac{1}{2\pi^{1/2}}\int_0^{\frac{bct}{a^{1/2}}} \tau\left(t - \frac{a^{1/2}\tau}{bc}\right)^{-3/2} \right. \\ &\qquad \left. \times \exp\left[-a\xi'^2\left(t - \frac{a^{1/2}\tau}{bc}\right) - \frac{\tau^2}{4(t - a^{1/2}\tau/bc)}\right] d\tau\right) \\ &= \frac{1}{4\pi^{1/2}}\frac{bc}{a^{3/2}}\int_0^t \tau(t-\tau)^{-5/2}\exp\left[-\frac{|z'|^2}{4a(t-\tau)} - \frac{c^2b^2\tau^2}{4a(t-\tau)}\right] d\tau \\ &= \frac{1}{4\pi^{1/2}}\frac{bc}{a^{3/2}}\int_0^t (t-\tau)\tau^{-5/2}\exp\left[-\frac{|z'|^2}{4a\tau} - \frac{c^2b^2(t-\tau)^2}{4a\tau}\right] d\tau, \end{aligned} \tag{3.13}$$

$\sigma(z',t)$ can be represented as

$$\sigma(z',t) = \int_0^t \int_{\mathbf{R}^2} K(z'-\xi', t-\tau) f(\xi',\tau) d\xi' d\tau.$$

From (3.13) it follows that

$$\left|D_t^r D_{z_1}^{s_1} D_{z_2}^{s_2} K(z',t)\right| \le C(t)\left(|z'|^2 + t^2\right)^{-1-(r+s_1+s_2)/2}.$$

Hence we have

$$\|\sigma\|_{R^2_\infty}^{(3+\alpha,\frac{3+\alpha}{2})} \le C\,|f|_{R^2_\infty}^{(2+\alpha,\frac{2+\alpha}{2})}. \tag{3.14}$$

On the other hand, in (3.11) w' can be considered as a solution of the Dirichlet problem of heat equation; we find the estimate for w' as

$$|w'|_{D^3_\infty}^{(3+\alpha,\frac{3+\alpha}{2})} \le Cb\,\|\sigma\|_{R^2_\infty}^{(3+\alpha,\frac{3+\alpha}{2})} \le C\,|f|_{R^2_\infty}^{(2+\alpha,\frac{2+\alpha}{2})}.$$

The nonhomogeneous case can be treated by adding the above w' to a solution w'' of the problem

$$\begin{cases} \dfrac{\partial w''}{\partial t} - a\Delta w'' = h \quad \text{in} \quad D^3_\infty, \\ w''\,\big|_{t=0} = 0 \quad \text{on} \quad R^3_+, \\ w'' = g \quad \text{on} \quad R^2_\infty. \end{cases}$$

Thus we have the following theorem.

THEOREM 3.2. *Let us assume*

$$F_5 \in C_0^{1+\alpha,\frac{1+\alpha}{2}}(Q_T),\ F_6 \in C_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T),\ F_7 \in C_0^{2+\alpha,\frac{2+\alpha}{2}}(\Gamma_T),\ G_2 \in C_0^{3+\alpha,\frac{3+\alpha}{2}}(\Sigma_T).$$

Then problem (3.2) *has a unique solution* $w \in C_0^{3+\alpha,\frac{3+\alpha}{2}}(Q_T), \sigma \in \tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T)$ *satisfying*

$$|w|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})} + \|\sigma\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})} \tag{3.15}$$

$$\le C\left(|F_5|_{Q_T}^{(1+\alpha,\frac{1+\alpha}{2})} + |F_6|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})} + |F_7|_{\Gamma_T}^{(2+\alpha,\frac{2+\alpha}{2})} + |G_2|_{\Sigma_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right),$$

where a constant C *depends only on* $F_j, j = 5, 6, 7, G_2$, *and* T *and remains bounded as* $T \to 0$.

Remark. Bazaliĭ and Degtyarev [3] derived the incorrect representation of the inverse Fourier–Laplace transformation of the symbol in (3.12); the correct formula is (3.13). However, the regularity of σ in the form of (3.14) can be obtained in the same way as their work.

4. Proof of Theorem 1.1. We begin with the estimates of $\mathcal{F}_j, j = 1, \dots, 6$ in (2.3) and (2.4).

LEMMA 4.1. *Let* $\sigma \in \tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T)$, $w \in C_0^{3+\alpha,\frac{3+\alpha}{2}}(Q_T)$ *and* $0 < \alpha' < \alpha < 1$. *Then the following inequalities hold for any* $\boldsymbol{u}_1, \boldsymbol{u}_2 \in C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T)$, $\nabla q_1, \nabla q_2 \in C_0^{\alpha,\frac{\alpha}{2}}(Q_T)$*:*

$$\begin{cases} |\mathcal{F}_1(\boldsymbol{u}_1, \nabla q_1, w, \sigma) - \mathcal{F}_1(\boldsymbol{u}_2, \nabla q_2, w, \sigma)|_{Q_T}^{(\alpha',\frac{\alpha'}{2})} \\ \qquad \le CT^{\frac{\alpha-\alpha'}{2}}\left(|\boldsymbol{u}_1 - \boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})} + |\nabla q_1 - \nabla q_2|_{Q_T}^{(\alpha,\frac{\alpha}{2})}\right), \\ |\mathcal{F}_2(\boldsymbol{u}_1, \sigma) - \mathcal{F}_2(\boldsymbol{u}_2, \sigma)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})} \le CT^{\frac{\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1 - \boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}, \\ |\mathcal{F}_3(\boldsymbol{u}_1, \sigma) - \mathcal{F}_3(\boldsymbol{u}_2, \sigma)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})} \le CT^{\frac{\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1 - \boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}, \\ |\mathcal{F}_4(\boldsymbol{u}_1, \sigma) - \mathcal{F}_4(\boldsymbol{u}_2, \sigma)|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})} \le CT^{\frac{\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1 - \boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}. \end{cases} \tag{4.1}$$

Let $\boldsymbol{u} \in C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T)$ *and* $\nabla q \in C_0^{\alpha,\frac{\alpha}{2}}(Q_T)$. *Then the following inequalities hold for any* $w_1, w_2 \in C_0^{3+\alpha,\frac{3+\alpha}{2}}(Q_T), \sigma_1, \sigma_2 \in \tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T)$*:*

$$
\text{(4.2)}\quad \begin{cases}
|\mathcal{F}_1(\boldsymbol{u},\nabla q,w_1,\sigma_1)-\mathcal{F}_1(\boldsymbol{u},\nabla q,w_2,\sigma_2)|_{Q_T}^{(\alpha',\frac{1+\alpha'}{2})} \\
\qquad \le CT^{\frac{2+\alpha-\alpha'}{2}}\left(\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}+|u_1-u_2|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right), \\
|\mathcal{F}_2(\boldsymbol{u},\sigma_1)-\mathcal{F}_2(\boldsymbol{u},\sigma_2)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})} \le CT^{\frac{2+\alpha-\alpha'}{2}}\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}, \\
|\mathcal{F}_3(\boldsymbol{u},\sigma_1)-\mathcal{F}_3(\boldsymbol{u},\sigma_2)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})} \le CT^{\frac{\alpha-\alpha'}{2}}\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}, \\
|\mathcal{F}_4(\boldsymbol{u},\sigma_1)-\mathcal{F}_4(\boldsymbol{u},\sigma_2)|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})} \le CT^{\frac{1+\alpha-\alpha'}{2}}\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}.
\end{cases}
$$

Here C*'s in* (4.1)–(4.2) *are constants depending on* $T, w, \sigma, \boldsymbol{u}, q$*, and bounded as* $T \to 0$.

Proof. Note that the definition of E_d implies

$$
\text{(4.3)}\quad \left|(E^*_{\sigma_1+\hat{d}})^{-1}-(E^*_{\sigma_2+\hat{d}})^{-1}\right|_{Q_T}^{(l+\alpha,\frac{l+\alpha}{2})} \le C\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(l+1+\alpha,\frac{l+1+\alpha}{2})}, \quad l=0,1,2.
$$

Then we have the following estimates:

$$
\begin{aligned}
&|\mathcal{F}_1(\boldsymbol{u}_1,\nabla q_1,w,\sigma)-\mathcal{F}_1(\boldsymbol{u}_2,\nabla q_2,w,\sigma)|_{Q_T}^{(\alpha',\frac{\alpha'}{2})} \\
&\quad=\Big|-(\boldsymbol{h}_{\sigma+\hat{d}}\cdot\nabla)(\boldsymbol{u}_1-\boldsymbol{u}_2)+\nu(\nabla^2_{\sigma+\hat{d}}-\nabla^2)(\boldsymbol{u}_1-\boldsymbol{u}_2) \\
&\qquad-\left((\boldsymbol{u}_1-\boldsymbol{u}_2)\cdot\nabla_{\sigma+\hat{d}}\right)(\boldsymbol{u}_1+\hat{\boldsymbol{v}})+\left((\boldsymbol{u}_2+\hat{\boldsymbol{v}})\cdot\nabla_{\sigma+\hat{d}}\right)(\boldsymbol{u}_1-\boldsymbol{u}_2) \\
&\qquad-(\nabla_{\sigma+\hat{d}}-\nabla)(q_1-q_2)\Big|_{Q_T}^{(\alpha',\frac{\alpha'}{2})} \\
&\quad\le C\left(|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha',\frac{2+\alpha'}{2})}+|\nabla q_1-\nabla q_2|_{Q_T}^{(\alpha',\frac{\alpha'}{2})}\right) \\
&\quad\le CT^{\frac{\alpha-\alpha'}{2}}\left(|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}+|\nabla q_1-\nabla q_2|_{Q_T}^{(\alpha,\frac{\alpha}{2})}\right),
\end{aligned}
$$

$$
\begin{aligned}
|\mathcal{F}_2(\boldsymbol{u}_1,\sigma)-\mathcal{F}_2(\boldsymbol{u}_2,\sigma)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})} &= \left|-(\nabla_{\sigma+\hat{d}}-\nabla)\cdot(\boldsymbol{u}_1-\boldsymbol{u}_2)\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})} \\
&\le C\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha',\frac{2+\alpha'}{2})} \\
&\le CT^{\frac{\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})},
\end{aligned}
$$

$$
\begin{aligned}
|\mathcal{F}_3(\boldsymbol{u}_1,\sigma)-\mathcal{F}_3(\boldsymbol{u}_2,\sigma)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})} &= \left|-(\boldsymbol{u}_1-\boldsymbol{u}_2)\cdot(\boldsymbol{n}_{\sigma+\hat{d}}-\boldsymbol{n})\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})} \\
&\le CT^{\frac{\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})},
\end{aligned}
$$

$$
\begin{aligned}
&|\mathcal{F}_4(\boldsymbol{u}_1,\sigma)-\mathcal{F}_4(\boldsymbol{u}_2,\sigma)|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})} \\
&\quad=\Big|-2\nu\Pi_{\sigma+\hat{d}}\mathbf{D}_{\sigma+\hat{d}}(\boldsymbol{u}_1-\boldsymbol{u}_2)\boldsymbol{n}_{\sigma+\hat{d}}+2\nu\Pi\mathbf{D}(\boldsymbol{u}_1-\boldsymbol{u}_2)\boldsymbol{n}
\end{aligned}
$$

$$
\begin{aligned}
&+\Pi_{\sigma+\hat{d}}\left[(\boldsymbol{u}_1-\boldsymbol{u}_2)\left(\boldsymbol{u}_1+\hat{\boldsymbol{v}}-\frac{\partial}{\partial t}(\sigma+\hat{d})\boldsymbol{n}_{\sigma+\hat{d}}\right)^*\right]\boldsymbol{n}_{\sigma+\hat{d}}\\
&+\Pi_{\sigma+\hat{d}}\left[(\boldsymbol{u}_2+\hat{\boldsymbol{v}})(\boldsymbol{u}_1-\boldsymbol{u}_2)^*\right]\boldsymbol{n}_{\sigma+\hat{d}}\Big|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\le C\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\le CT^{\frac{\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}.
\end{aligned}
$$

Further we have the following estimates:

$$
\begin{aligned}
&|\mathcal{F}_1(\boldsymbol{u},\nabla q,w_1,\sigma_1)-\mathcal{F}_1(\boldsymbol{u},\nabla q,w_2,\sigma_2)|_{Q_T}^{(\alpha',\frac{\alpha'}{2})}\\
&\quad=\Bigg|\boldsymbol{f}\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)-\boldsymbol{f}\left(w_2+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\\
&\qquad-\left((\boldsymbol{h}_{\sigma_1+\hat{d}}-\boldsymbol{h}_{\sigma_2+\hat{d}})\cdot\nabla\right)(\boldsymbol{u}+\hat{\boldsymbol{v}})+\nu\left(\nabla^2_{\sigma_1+\hat{d}}-\nabla^2_{\sigma_2+\hat{d}}\right)(\boldsymbol{u}+\hat{\boldsymbol{v}})\\
&\qquad-\left((\boldsymbol{u}+\hat{\boldsymbol{v}})\cdot(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}})\right)(\boldsymbol{u}+\hat{\boldsymbol{v}})-(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}})(q+\hat{p})\Bigg|_{Q_T}^{(\alpha',\frac{\alpha'}{2})}\\
&\quad\le C|D\boldsymbol{f}|^{(\alpha)}\left(|w_1-w_2|_{Q_T}^{(\alpha',\frac{\alpha'}{2})}+|\sigma_1-\sigma_2|_{Q_T}^{(\alpha',\frac{\alpha'}{2})}\right)\\
&\qquad+C\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad\le CT^{\frac{2+\alpha-\alpha'}{2}}\left(\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}+|w_1-w_2|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right),
\end{aligned}
$$

$$
\begin{aligned}
&|\mathcal{F}_2(\boldsymbol{u},\sigma_1)-\mathcal{F}_2(\boldsymbol{u},\sigma_2)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad=\left|-(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}})\cdot\boldsymbol{u}-(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}})\cdot\hat{\boldsymbol{v}}\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad\le C\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad\le CT^{\frac{2+\alpha-\alpha'}{2}}\,\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})},
\end{aligned}
$$

$$
\begin{aligned}
&|\mathcal{F}_3(\boldsymbol{u},\sigma_1)-\mathcal{F}_3(\boldsymbol{u},\sigma_2)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\quad=\Bigg|-\boldsymbol{u}\cdot\left(\boldsymbol{n}_{\sigma_1+\hat{d}}-\boldsymbol{n}_{\sigma_2+\hat{d}}\right)-\hat{\boldsymbol{v}}\cdot\left(\boldsymbol{n}_{\sigma_1+\hat{d}}-\boldsymbol{n}_{\sigma_2+\hat{d}}\right)\\
&\qquad-\left(1-\frac{\varrho_e}{\varrho}\right)\left(\frac{\partial\sigma_1}{\partial t}-\frac{\partial\sigma_2}{\partial t}\right)\Bigg|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\quad\le C\left(\left|\boldsymbol{n}_{\sigma_1+\hat{d}}-\boldsymbol{n}_{\sigma_2+\hat{d}}\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}+\left|\frac{\partial\sigma_1}{\partial t}-\frac{\partial\sigma_2}{\partial t}\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\right)\\
&\quad\le C\left(\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}+\left|\frac{\partial\sigma_1}{\partial t}-\frac{\partial\sigma_2}{\partial t}\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\right)\\
&\quad\le CT^{\frac{\alpha-\alpha'}{2}}\,\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})},
\end{aligned}
$$

$$
\begin{aligned}
&|\mathcal{F}_4(\boldsymbol{u},\sigma_1)-\mathcal{F}_4(\boldsymbol{u},\sigma_2)|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad=\Bigg|-2\nu\Pi_{\sigma_1+\hat{d}}\mathbf{D}_{\sigma_1+\hat{d}}(\boldsymbol{u}+\hat{\boldsymbol{v}})\boldsymbol{n}_{\sigma_1+\hat{d}}+2\nu\Pi_{\sigma_2+\hat{d}}\mathbf{D}_{\sigma_2+\hat{d}}(\boldsymbol{u}+\hat{\boldsymbol{v}})\boldsymbol{n}_{\sigma_2+\hat{d}}\\
&\qquad+\Pi_{\sigma_1+\hat{d}}\left[(\boldsymbol{u}+\hat{\boldsymbol{v}})\left(\boldsymbol{u}+\hat{\boldsymbol{v}}-\frac{\partial}{\partial t}(\sigma_1+\hat{d})\boldsymbol{n}_{\sigma_1+\hat{d}}\right)^*\right]\boldsymbol{n}_{\sigma_1+\hat{d}}\\
&\qquad-\Pi_{\sigma_2+\hat{d}}\left[(\boldsymbol{u}+\hat{\boldsymbol{v}})\left(\boldsymbol{u}+\hat{\boldsymbol{v}}-\frac{\partial}{\partial t}(\sigma_2+\hat{d})\boldsymbol{n}_{\sigma_2+\hat{d}}\right)^*\right]\boldsymbol{n}_{\sigma_2+\hat{d}}\Bigg|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad\le C\left(\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})}+\left|\frac{\partial\sigma_1}{\partial t}-\frac{\partial\sigma_2}{\partial t}\right|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\right)\\
&\quad\le CT^{\frac{1+\alpha-\alpha'}{2}}\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}.
\end{aligned}
$$

Hence the proof is completed.

LEMMA 4.2. *Let $\boldsymbol{u}\in C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T)$. Then the following inequalities hold for any $w_1,w_2\in C_0^{3+\alpha,\frac{3+\alpha}{2}}(Q_T)$, $\sigma_1,\sigma_2\in\tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T)$:*

$$
\left\{\begin{aligned}
&|\mathcal{F}_5(\boldsymbol{u},w_1,\sigma_1)-\mathcal{F}_5(\boldsymbol{u},w_2,\sigma_2)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad\le CT^{\frac{\alpha-\alpha'}{2}}\left(|u_1-u_2|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})}+\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right),\\
&|\mathcal{F}_6(w_1,\sigma_1)-\mathcal{F}_6(w_2,\sigma_2)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\quad\le CT^{\frac{\alpha-\alpha'}{2}}\left(|w_1-w_2|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})}+\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right).
\end{aligned}\right.\tag{4.4}
$$

If $w\in C_0^{3+\alpha,\frac{3+\alpha}{2}}(Q_T),\sigma\in\tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T)$, then the following inequality holds for any $\boldsymbol{u}_1,\boldsymbol{u}_2\in C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T)$.

$$
|\mathcal{F}_5(\boldsymbol{u}_1,w,\sigma)-\mathcal{F}_5(\boldsymbol{u}_2,w,\sigma)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\le CT^{\frac{1+\alpha-\alpha'}{2}}|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}.\tag{4.5}
$$

Here C's in (4.4) *and* (4.5) *are constants depending on $T,w,\sigma,\boldsymbol{u},q$, and bounded as $T\to 0$.*

Proof. In the same way as in the proof of Lemma 4.1, we can prove (4.4) and (4.5). Here we give only the estimates of the most complicated terms in each of $\mathcal{F}_5$ and $\mathcal{F}_6$. For $(4.4)^1$,

$$
\begin{aligned}
&\Bigg|\nabla_{\sigma_1+\hat{d}}\cdot\left(\kappa\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\nabla_{\sigma_1+\hat{d}}\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right)\\
&\qquad-\nabla_{\sigma_2+\hat{d}}\cdot\left(\kappa\left(w_2++\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\nabla_{\sigma_2+\hat{d}}\left(w_2+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\right)\Bigg|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\le\Bigg|\left(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}}\right)\cdot\left(\kappa\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right.\\
&\qquad\left.\nabla_{\sigma_1+\hat{d}}\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right)\Bigg|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}
\end{aligned}
$$

$$
\begin{aligned}
&+\left|\nabla_{\sigma_2+\hat{d}}\cdot\left(\left(\kappa\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)-\kappa\left(w_2+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\right)\right.\right.\\
&\qquad\left.\left.\cdot\nabla_{\sigma_1+\hat{d}}\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right)\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&+\left|\nabla_{\sigma_2+\hat{d}}\cdot\left(\kappa\left(w_2+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\right.\right.\\
&\qquad\left.\left.\cdot\left(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}}\right)\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right)\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&+\left|\nabla_{\sigma_2+\hat{d}}\cdot\left(\kappa\left(w_2+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\right.\right.\\
&\qquad\left.\left.\cdot\nabla_{\sigma_2+\hat{d}}\left(w_1-w_2+\chi\frac{\partial\hat{\theta}}{\partial n}(\sigma_1-\sigma_2)\right)\right)\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\le C\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad+C\left|D^2\kappa\right|^{(\alpha)}\left|w_1-w_2+\chi\frac{\partial\hat{\theta}}{\partial n}(\sigma_1-\sigma_2)\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad+C\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{Q_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\quad+C\left|w_1-w_2+\chi\frac{\partial\hat{\theta}}{\partial n}(\sigma_1-\sigma_2)\right|_{Q_T}^{(3+\alpha',\frac{3+\alpha'}{2})}\\
&\le CT^{\frac{\alpha-\alpha'}{2}}\left(|w_1-w_2|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})}+\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right).
\end{aligned}
$$

For $(4.4)^2$,

$$
\begin{aligned}
&\left|\left(\nabla_{\sigma_1+\hat{d}}\eta\cdot\nabla_{\sigma_1+\hat{d}}\right)\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right.\\
&\qquad\left.-\left(\nabla_{\sigma_2+\hat{d}}\eta\cdot\nabla_{\sigma_2+\hat{d}}\right)\left(w_2+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_2\right)\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\le C\left|\left(\left(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}}\right)\eta\cdot\nabla_{\sigma_1+\hat{d}}\right)\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\quad+C\left|\left(\nabla_{\sigma_2+\hat{d}}\eta\cdot\left(\nabla_{\sigma_1+\hat{d}}-\nabla_{\sigma_2+\hat{d}}\right)\right)\left(w_1+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma_1\right)\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}
\end{aligned}
$$

$$
\begin{aligned}
&+C\left|\left(\nabla_{\sigma_2+\hat{d}}\eta\cdot\nabla_{\sigma_2+\hat{d}}\right)\left(w_1-w_2+\chi\frac{\partial\hat{\theta}}{\partial n}(\sigma_1-\sigma_2)\right)\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\le C\left|\left(E^*_{\sigma_1+\hat{d}}\right)^{-1}-\left(E^*_{\sigma_2+\hat{d}}\right)^{-1}\right|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})}\\
&\quad+C\left|w_1-w_2+\chi\frac{\partial\hat{\theta}}{\partial n}(\sigma_1-\sigma_2)\right|_{\Gamma_T}^{(3+\alpha',\frac{3+\alpha'}{2})}\\
&\le CT^{\frac{\alpha-\alpha'}{2}}\left(|w_1-w_2|_{Q_T}^{(3+\alpha,\frac{3+\alpha}{2})}+\|\sigma_1-\sigma_2\|_{\Gamma_T}^{(3+\alpha,\frac{3+\alpha}{2})}\right).
\end{aligned}
$$

And for (4.5),

$$
\begin{aligned}
&\left|(\boldsymbol{u}_1\cdot\nabla_{\sigma+\hat{d}})\left(w+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma\right)-(\boldsymbol{u}_2\cdot\nabla_{\sigma+\hat{d}})\left(w+\hat{\theta}+\chi\frac{\partial\hat{\theta}}{\partial n}\sigma\right)\right|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\\
&\quad\le C\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})}\le CT^{\frac{1+\alpha-\alpha'}{2}}\,|\boldsymbol{u}_1-\boldsymbol{u}_2|_{Q_T}^{(2+\alpha,\frac{2+\alpha}{2})}.
\end{aligned}
$$

Hence the proof is completed.

Now we proceed to the proof of Theorem 1.1. Set

$$
X\equiv\left\{\begin{array}{l}(w,\sigma)\in C_0^{3+\alpha,\frac{3+\alpha}{2}}(Q_T)\times\tilde{C}_0^{3+\alpha,\frac{3+\alpha}{2}}(\Gamma_T)\;\big|\\[2ex] \|(w,\sigma)\|_X\equiv|w|_{Q_T}^{(3+\alpha',\frac{3+\alpha'}{2})}+\|\sigma\|_{\Gamma_T}^{(3+\alpha',\frac{3+\alpha}{2})}\le C\end{array}\right\},
$$

$$
Y\equiv\left\{\begin{array}{l}(\boldsymbol{u},\nabla q)\in C_0^{2+\alpha,\frac{2+\alpha}{2}}(Q_T)\times C_0^{\alpha,\frac{\alpha}{2}}(Q_T)\;\big|\\[2ex] \|(\boldsymbol{u},\nabla q)\|_Y\equiv|\boldsymbol{u}|_{Q_T}^{(2+\alpha',\frac{2+\alpha'}{2})}+|\nabla q|_{Q_T}^{(\alpha',\frac{\alpha'}{2})}\le C\end{array}\right\},
$$

where C is a positive constant. And let operators P and Q,

$$
P:((w_1,\sigma_1),(\boldsymbol{u}_1,\nabla q_1))\longmapsto((w_1,\sigma_1),(\boldsymbol{u}_2,\nabla q_2)),
$$

$$
Q:((w_1,\sigma_1),(\boldsymbol{u}_1,\nabla q_1))\longmapsto((w_2,\sigma_2),(\boldsymbol{u}_1,\nabla q_1)),
$$

assign the solution $(\boldsymbol{u}_2,\nabla q_2)$ of problem (2.3) and the solution (w_2,σ_2) of problem (2.4) for arbitrary given $((w_1,\sigma_1),(\boldsymbol{u}_1,\nabla q_1))$ in $X\times Y$, respectively. By using Lemmas 4.1 and 4.2, first it is easily seen that the operator $Q\circ P$ is continuous, i.e.,

$$
\begin{aligned}
&\|Q\circ P((w,\sigma),(\boldsymbol{u},\nabla q))-Q\circ P((w',\sigma'),(\boldsymbol{u}',\nabla q'))\|_{X\times Y}\\
&\quad\le C\|((w,\sigma),(\boldsymbol{u},\nabla q))-((w',\sigma'),(\boldsymbol{u}',\nabla q'))\|_{X\times Y}
\end{aligned}
$$

for arbitrary $((w,\sigma),(\boldsymbol{u},\nabla q)),((w',\sigma'),(\boldsymbol{u}',\nabla q'))$ in $X\times Y$, and some constant C.

Next for arbitrary $((w,\sigma),(\boldsymbol{u},\nabla q))$ in $X\times Y$ we have

$$
\begin{aligned}
&\|Q\circ P((w,\sigma),(\boldsymbol{u},\nabla q))\|_{X\times Y}\\
&\quad\le\|Q\circ P((w,\sigma),(\boldsymbol{u},\nabla q))-Q\circ P((0,0),(0,0))\|_{X\times Y}
\end{aligned}
$$

$$
\begin{aligned}
&+ \|Q \circ P((0,0),(0,0))\|_{X\times Y} \\
&\le C\,\|(w,\sigma),(\boldsymbol{u},\nabla q))\|_{X\times Y} \\
&+ |\mathcal{F}_1(0,0,0,0)|_{Q_T}^{(\alpha',\frac{\alpha'}{2})} + |\mathcal{F}_2(0,0)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})} + |\mathcal{F}_3(0,0)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})} \\
&+ |\mathcal{F}_4(0,0)|_{\Gamma_T}^{(1+\alpha',\frac{1+\alpha'}{2})} + |\mathcal{F}_5(0,0,0)|_{Q_T}^{(1+\alpha',\frac{1+\alpha'}{2})} + |\mathcal{F}_6(0,0)|_{\Gamma_T}^{(2+\alpha',\frac{2+\alpha'}{2})} \\
&\le C(T),
\end{aligned}
$$

where $C(T)$ is a constant which tends to 0 as $T \to 0$. Here we have used the fact that $((w,\sigma),(\boldsymbol{u},\nabla q))$ and $\mathcal{F}_i, i = 1,\dots,6$ are equal to zero at $t = 0$ in the higher-order classes. Hence $Q \circ P$ maps $X \times Y$ into itself for small T. Therefore Shauder's fixed point theorem yields the solution of problem (2.2).

REFERENCES

[1] S. Agmon, S. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] B. V. Bazaliĭ and S. P. Degtyarev, *The classical Stefan problem as the limit case of the Stefan problem with a kinetic condition at the free boundary*, in Free Boundary Problems in Continuum Mechanics, S. N. Antontsev, K. H. Hoffmann, and A. M. Khludnev, eds., Birkhäuser Verlag, Basel, 1992, pp. 83–90.

[3] B. V. Bazaliĭ and S. P. Degtyarev, *On classical solvability of the multidimensional Stefan problem for convective motion of a viscous incompressible fluid*, Mat. Sb. (N.S.), 132 (1987), pp. 3–19; English transl. in Math. USSR Sb., 60 (1988), pp. 1–17.

[4] J. A. Bilenas and L. M. Jiji, *Variational solution of axisymmetric fluid flow in tubes with surface solidification,* J. Franklin Inst., 289 (1970), pp. 265–279.

[5] G. Caginalp, *Stefan and Hele-Shaw type model as asymptotic limits of the phase field equations*, Phys. Rev., 39 (1989), pp. 5887–5896.

[6] X. Chen and F. Reitich, *Local existence and uniqueness of the Stefan problem with kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.

[7] *Chronological Scientific Tables*, edited by National Astronomical Observatory, Maruzen, Tokyo, 1997.

[8] D. G. Donald and G. Aldo, *The validity of the Boussinesq approximation for liquids and gases*, Int. J. Heat Mass Transfer., 19 (1975), pp. 545–551.

[9] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice Hall, Englewood Cliffs., NJ, 1964.

[10] E. Hanzawa, *Classical solutions of the Stefan problem*, Tohoku Math. J., (2) 33 (1981), pp. 297–335.

[11] M. Ishii, *Thermo-fluid Dynamic Theory of Two-Phase Flow*, Eyrolles, Paris, France, 1975.

[12] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[13] L. D. Landau and E. M. Lifshitz, *Fluid Mechanics*, Pergamon Press, Oxford, UK, Elmsford, NY, 1963.

[14] A. M. Meĭrmanov, *On the classical solution of the multidimensional Stefan problem for quasilinear parabolic equations*, Mat. Sb. (N.S.), 112 (1980), pp. 170–192; English transl. in Math. USSR Sb., 40 (1981), pp. 157–178.

[15] A. M. Meĭrmanov, *The Stefan Problem*, Walter de Gruyter, Berlin, New York, 1992.

[16] E. Michael and F. B. Cheung, *Complex freezing-melting interfaces in fluid flow*, Ann. Rev. Fluid Mech., 15 (1983), pp. 293–319.

[17] S. Morioka and R. Ishii, *Fundamental Equations*, in Fluid Dynamics of Multiphase Flow, S. Morioka, ed., Asakura Shoten, Tokyo, 1991 (in Japanese).

[18] E. V. Radkevich, *On conditions for existence of a classical solution of the modified Stefan problem (the Gibbs-Thomson law)*, Russian Acad. Sci. Sb. Math., 75 (1993), pp. 221–246.

[19] L. I. Rubinstein, *The Stefan Problem*, Transl. Math. Monogr. 27, AMS, Providence, RI, 1971.
[20] V. A. Solonnikov, *Estimates of solutions of an initial and boundary value problem for the linear nonstationary Navier-Stokes system*, Zap. Nauchn. Sem. LOMI, 59 (1976), pp. 336–393; English transl. in J. Soviet Math., 10 (1978), pp. 178–254.
[21] V. A. Solonnikov, *Estimates for solutions of nonstationary Navier-Stokes equations*, Zap. Nauchn, Sem. LOMI, 38 (1973), pp. 153–231; English transl. in J. Soviet Math., 8 (1977), pp. 467–529.
[22] Y. Tao, *Classical solution of Verigin problem with surface tension*, CNS preprint series 025, Suzhou University, Saochow, China, 1994.
[23] D. Takahashi and H. Takami, *Numerical simulation of fluid flow with solidification*, in Solutions of the Navier-Stokes Equations, H. Takami, ed., RIMS Kokyuroku 539, RIMS Kyoto University, Kyoto, Japan, 1984, pp. 194–214 (in Japanese).
[24] A. Tani, S. Itoh, and N. Tanaka, *The initial value problem for the Navier-Stokes equations with general slip boundary condition*, Adv. Math. Sci. Appl., 4 (1994), pp. 51–69.
[25] M. Yamaguti and T. Nogi, *The Stefan Problem*, Sangyo Tosho, Tokyo, 1977 (in Japanese).

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 603–622

A VERTICAL DIFFUSION MODEL FOR LAKES*

DIDIER BRESCH†, JÉRÔME LEMOINE†, AND JACQUES SIMON†

Abstract.
The motion of a fluid in a lake with small depth compared to width is investigated. We prove that when the depth goes to 0, the solution of the stationary Navier–Stokes equations with adherence at the bottom and traction by wind at the surface, once conveniently normalized, goes to a three-dimensional limit which is the solution of an incompressible model with vertical diffusion.

The limit velocity is given in terms of the vertical coordinate and of the limit pressure. This pressure, which depends only on the horizontal coordinates, is driven by a two-dimensional equation on the surface degenerating on the shore, which is solved in a weighted space. Thus, a three-dimensional approximation is obtained by a simple two-dimensional computation.

Key words. Navier–Stokes equations, thin domains, asymptotic analysis, hydrostatic

AMS subject classifications. 35Q30, 35B40, 76D05

PII. S0036141097322947

Introduction. *A simplified model for lakes.* The fluid occupies the following domain in $\mathbb{R}^3$

$$d = \{(x,z) : x \in \Gamma,\ -h(x) < z < 0\},$$

where Γ is an open bounded set in $\mathbb{R}^2$ (the horizontal section) and where the depth h is a positive continuous function on Γ, vanishing on $\partial\Gamma$. The boundary is $\partial d = f \cup s \cup \partial s$ (see Figure 1), where the bottom f, the surface s, and the shore ∂s are

$$f = \{(x,-h(x)) : x \in \Gamma\}, \quad s = \{(x,0) : x \in \Gamma\}, \quad \partial s = \partial f = \{(x,0) : x \in \partial\Gamma\}.$$

The horizontal velocity $v = (v_1, v_2)$, the vertical velocity w, and the pressure p satisfy the stationary Navier–Stokes equations in d, which may be written as

$$\begin{cases} -\nu(\Delta v + \partial_z^2 v) + kBv + k'we + (v\cdot\nabla)\,v + w\,\partial_z v + \nabla p = 0, \\ -\nu(\Delta w + \partial_z^2 w) - k'v_1 + (v\cdot\nabla)\,w + w\,\partial_z w + \partial_z p = 0, \\ \nabla\cdot v + \partial_z w = 0, \end{cases} \tag{1}$$

where $\nabla = (\partial/\partial x_1, \partial/\partial x_2)$, $\Delta = \partial^2/\partial x_1^2 + \partial^2/\partial x_2^2$, $\partial_z = \partial/\partial z$, $B(v_1,v_2) = (-v_2, v_1)$, $e = (1,0)$, $k = |\omega|\sin\theta$, $k' = |\omega|\cos\theta$, ω being the rotating velocity of the earth and θ the latitude, and $\nu > 0$. This system is completed by the adherence on the bottom and by the horizontal traction at the surface:

$$v = 0,\ w = 0 \text{ on } f, \quad \nu\partial_z v = g,\ w = 0 \text{ on } s, \tag{2}$$

where the force due to the wind $g = (g_1, g_2)$ is a given function on s.

*Received by the editors June 13, 1997; accepted for publication May 27, 1998; published electronically April 7, 1999.
http://www.siam.org/journals/sima/30-3/32294.html
†Laboratoire de Mathématiques Appliquées, Université Blaise Pascal (Clermont-Ferrand 2), 63-177 Aubière Cedex, France (bresch@ucfma.univ-bpclermont.fr, lemoine@ucfma.univ-bpclermont.fr, simon@ucfma.univ-bpclermont.fr).

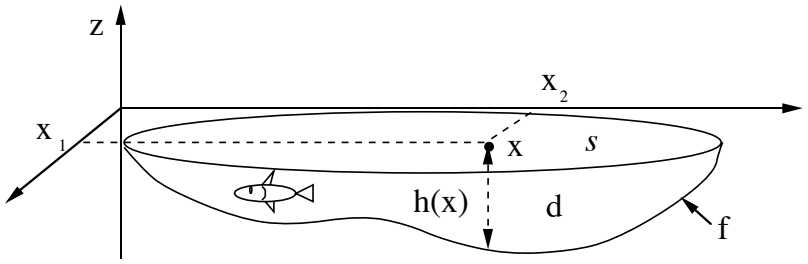


FIG. 1. *The domain occupied by the fluid.*

Asymptotic analysis. The aspect ratio ε of the domain, that is, the ratio of the height to the width, is assumed to be very small. Introducing a normalized depth H, we will investigate the behavior of the solution as

$$h^\varepsilon = \varepsilon H, \quad \varepsilon \to 0. \tag{3}$$

Denoting by $(u^\varepsilon, v^\varepsilon, w^\varepsilon)$ the solution of (1) and (2) and d^ε, s^ε, f^ε the domain, the surface, and the bottom associated with h^ε, we will prove that

$$\begin{cases} v^\varepsilon(x,z) \approx \dfrac{\varepsilon}{\nu}\left(\dfrac{1}{2}(r^2-1)H^2(x)\nabla\mathbf{P}(x) + (1-r)H(x)g(x)\right), \\ w^\varepsilon(x,z) \approx \dfrac{\varepsilon^2}{\nu}\nabla\cdot\left(\left(\dfrac{r^3}{6}-\dfrac{r}{2}\right)H^3(x)\nabla\mathbf{P}(x) + \left(r-\dfrac{r^2}{2}\right)H^2(x)g(x)\right), \\ p^\varepsilon(x,z) \approx \dfrac{1}{\varepsilon}\mathbf{P}(x), \end{cases} \tag{4}$$

where r is the relative depth in d^ε, which is given by $r = -z/(\varepsilon H(x))$ and thus

$$0 \le r \le 1, \quad r = 0 \text{ at the surface } s^\varepsilon, \quad r = 1 \text{ at the bottom } f^\varepsilon,$$

and where $\mathbf{P}$ is the solution (unique up to a constant) of the two-dimensional (2D) equation in Γ:

$$\nabla\cdot\left(\frac{1}{3}H^3\nabla\mathbf{P} - \frac{1}{2}H^2 g\right) = 0, \quad \left(\frac{1}{3}H^3\nabla\mathbf{P} - \frac{1}{2}H^2 g\right)\cdot n_{\partial\Gamma} = 0. \tag{5}$$

At first, we will prove (Theorem 4) that there exists a solution $(v^\varepsilon, w^\varepsilon, p^\varepsilon)$ of (1) and (2). Then we will prove (Theorem 7) that, after normalization, it converges to the solution of the incompressible model with vertical diffusion (10). Finally, we will prove (Theorem 10) that this solution is given by the right-hand side of (4) and by (5).

Interest of the results. We obtain three-dimensional (3D) information—vertical variations of the horizontal and vertical velocities—for the simple cost of a 2D computation. In addition, the approximation (4) is divergence free.

Numerical simulations [5] show that this approximation of the velocity gives a good feature of the driving by the wind at the surface, of the forward motion at the bottom, and of the horizontal and vertical deflection of the fluid by the submarine topography.

No regularity on the section $\partial\Gamma$ is assumed, excepted that d is star-shaped with respect to one point of s in Theorem 10, and we suppose only that the depth h is continuous on $\overline{\Gamma}$ and that the force g due to the wind satisfies $h^{1/2}g \in (L^2(\Gamma))^2$.

Comparison with previous results. For lubrication problems, a similar asymptotic analysis without Coriolis force has been done in [1], [2]. It gives a Reynolds equation on $\mathbf{P}$ similar to (5) but which does not degenerate on the shore, since the depth possesses a positive lower bound (the distance between pieces) which avoids the use of weighted spaces as here.

Many authors have studied nonstationary problems in thin domains with a constant depth (see [8], [12], and their references), which do not describe our problem, for which the topography is essential. The results presented in this paper have been announced in [6]. An outline of these results is listed here.

1. Formal calculations.

1.1. Notations. We distinguish the horizontal variables (2D) from the vertical ones (1D), which play a particular part: $x = (x_1, x_2)$ are the horizontal coordinates, z is the vertical one; $v = (v_1, v_2)$ is the horizontal component of the velocity, w its vertical component.

We denote by small letters the quantities in the varying domain: (x, z), the coordinates; h^ε, the depth; d^ε, the domain; f^ε, s^ε, its bottom and its surface; g, the force due to the wind; and $(v^\varepsilon, w^\varepsilon, p^\varepsilon)$, the velocity and the pressure.

We denote by capital letters the scaled quantities in the fixed domain after change of variable: (X, Z), the coordinates; H, the depth; D, the domain; F, S, its bottom and its surface; G, the force due to the wind; $(V^\varepsilon, W^\varepsilon, P^\varepsilon)$, the scaled velocity and pressure.

We denote in boldface the limits as $\varepsilon \to 0$: $(\mathbf{V}, \mathbf{W}, \mathbf{P})$, the limit of the scaled velocity and pressure.

1.2. Derivation of the vertical diffusion model. We use the scaling

$$x = X, \quad z = \varepsilon Z, \tag{6}$$

to get a fixed domain in $\mathbb{R}^3$,

$$D = \{(X, Z) : X \in \Gamma,\ -H(X) < Z < 0\},$$

with boundary $\partial D = F \cup S \cup \partial S$. Its bottom is $F = \{(X, -H(X)) : X \in \Gamma\}$; the surface $S = s$ and the shore $\partial S = \partial s$ are unchanged. The force due to the wind, $G = g$, is unchanged. The transported velocity and pressure are scaled to be of magnitude one. More precisely, $(V^\varepsilon, W^\varepsilon, P^\varepsilon)$ is defined by

$$v^\varepsilon(x, z) = \frac{\varepsilon}{\nu} V^\varepsilon(X, Z), \quad w^\varepsilon(x, z) = \frac{\varepsilon^2}{\nu} W^\varepsilon(X, Z), \quad p^\varepsilon(x, z) = \frac{1}{\varepsilon} P^\varepsilon(X, Z). \tag{7}$$

Equations (1) and (2) scaled with (6) and (7) yield

$$
(8)\quad \begin{cases} -\partial_Z^2 V^\varepsilon + \nabla P^\varepsilon + \dfrac{\varepsilon^2 k}{\nu} B V^\varepsilon - \varepsilon^2 \Delta V^\varepsilon + \dfrac{\varepsilon^3 k'}{\nu} W^\varepsilon e \\ \qquad\qquad + \dfrac{\varepsilon^3}{\nu^2}\Big((V^\varepsilon \cdot \nabla) V^\varepsilon + W^\varepsilon \, \partial_Z V^\varepsilon \Big) = 0, \\ \partial_Z P^\varepsilon - \dfrac{\varepsilon^3 k'}{\nu} V_1^\varepsilon - \varepsilon^2 \partial_Z^2 W^\varepsilon - \varepsilon^4 \Delta W^\varepsilon + \dfrac{\varepsilon^5}{\nu^2}\Big((V^\varepsilon \cdot \nabla) W^\varepsilon + W^\varepsilon \, \partial_Z W^\varepsilon \Big) = 0, \\ \nabla \cdot V^\varepsilon + \partial_Z W^\varepsilon = 0 \text{ in } D; \end{cases}
$$

$$
(9)\qquad V^\varepsilon = 0,\ W^\varepsilon = 0 \text{ on } F; \quad \partial_Z V^\varepsilon = G,\ W^\varepsilon = 0 \text{ on } S.
$$

The possible limit $(\mathbf{V}, \mathbf{W}, \mathbf{P})$ of $(V^\varepsilon, W^\varepsilon, P^\varepsilon)$ as $\varepsilon \to 0$ must satisfy

$$
(10)\qquad \begin{cases} -\partial_Z^2 \mathbf{V} + \nabla \mathbf{P} = 0, \quad \partial_Z \mathbf{P} = 0, \quad \nabla \cdot \mathbf{V} + \partial_Z \mathbf{W} = 0 \quad \text{in } D; \\ \mathbf{V} = 0, \mathbf{W} = 0 \text{ on } F; \quad \partial_Z \mathbf{V} = G, \mathbf{W} = 0 \text{ on } S. \end{cases}
$$

We will see (section 1.4) that some information on the shore ∂S has been lost here.

Remark. A limit model with isotropic diffusion, namely $-\nu(\partial_Z^2 + \Delta)\mathbf{V} + \nabla \mathbf{P} = 0$, is obtained if the viscosity is assumed to be conveniently anisotropic—that is, if $\nu \Delta v$ is replaced in (1) by $\varepsilon^{-2} \nu \Delta v$; see [4]. □

1.3. Solution of the vertical diffusion model. Let us solve (10) with respect to Z for a fixed X. The second equation gives $\mathbf{P} = \mathbf{P}(X)$. The first equation and the conditions $\mathbf{V}(-H) = 0$ and $\partial_Z \mathbf{V}(0) = G$ give

$$
(11)\qquad \mathbf{V} = \frac{1}{2}(Z^2 - H^2)\nabla \mathbf{P} + (Z + H) G.
$$

The third equation and the condition $\mathbf{W}(0) = 0$ yield $\mathbf{W} = \int_Z^0 \nabla \cdot \mathbf{V} = \nabla \cdot (\int_Z^0 \mathbf{V})$ whence

$$
(12)\qquad \mathbf{W} = \nabla \cdot \left(\left(\frac{Z H^2}{2} - \frac{Z^3}{6} \right) \nabla \mathbf{P} - \left(\frac{1}{2} Z^2 + H Z \right) G \right).
$$

It remains to take into account the condition $\mathbf{W}(-H) = 0$. Since we have $\mathbf{W}(-H) = \int_{-H}^0 \nabla \cdot \mathbf{V} = \nabla \cdot (\int_{-H}^0 \mathbf{V}) - \nabla H \cdot \mathbf{V}(-H)$ and $\mathbf{V}(-H) = 0$, it gives $\nabla \cdot (\int_{-H}^0 \mathbf{V}) = 0$, that is,

$$
(13)\qquad \nabla \cdot \left(\frac{1}{3} H^3 \nabla \mathbf{P} - \frac{1}{2} H^2 G \right) = 0 \quad \text{in } \Gamma.
$$

Therefore (10) is formally equivalent to (11)–(13).

This provides the right-hand sides of the announced approximations (4) and the equation on $\mathbf{P}$ included in (5). The limit condition announced in (5) is missing, however.

1.4. The lost condition. Equation (13) does not define a unique $\mathbf{P}$ since there is no boundary condition on $\partial\Gamma$.

This comes from the fact that the boundary conditions (9) are assigned only on F and S but not on the shore ∂S (recall that $\partial D = F \cup S \cup \partial S$). On the shore,

$V^\varepsilon = 0$, but this condition does not pass to the limit. Indeed $\mathbf{V} = 0$ on ∂S would give contradictory results in the expression (11), whether the shore is considered to be at the surface (i.e., $Z = 0$) or at the bottom (i.e., $Z = -H$, which is equal to 0 too!). Besides this contradiction, it would not give the true answer (see Theorem 10), which is the Neumann condition included in (5).

Thus, we have to use a weaker condition than $V^\varepsilon|_{\partial S} = 0$, but one which goes to the limit. The condition $V^\varepsilon \cdot n_{\partial S} = 0$ does not fit, for the same reasons. We will use the fact that the flux through the shore cancels. The total flux on the vertical passing through X is $\overline{V^\varepsilon}(X) = \int_{-H(X)}^{0} V^\varepsilon dZ$. On ∂S, it vanishes since H and V^ε do; thus $\overline{V^\varepsilon} \cdot n_{\partial\Gamma} = 0$, where $n_{\partial\Gamma}$ is the normal to $\partial\Gamma$ in $\mathbb{R}^2$ (that is the horizontal normal to ∂S). At the limit,

$$\overline{\mathbf{V}} \cdot n_{\partial\Gamma} = 0, \tag{14}$$

that is, due to (11),

$$\left(\frac{1}{3}H^3\nabla\mathbf{P} - \frac{1}{2}H^2G\right) \cdot n_{\partial\Gamma} = 0, \tag{15}$$

which is the Neumann condition included in (5). It closes (13), ensuring the existence and uniqueness of $\mathbf{P}$ in the weighted space $\mathbb{A}(\Gamma)$ defined by (58).

Remark. We cannot write (15) as $\frac{1}{3}H^3\partial\mathbf{P}/\partial n - \frac{1}{2}H^2G \cdot n = 0$ since $\partial\mathbf{P}/\partial n$ in general has no sense. If it had some sense, it would be a truism since $H = 0$ on $\partial\Gamma$. We will give a sense to the left-hand side of (15), that is, to the normal trace $\overline{\mathbf{V}} \cdot n_{\partial\Gamma}$, by using $\nabla \cdot \overline{\mathbf{V}} = 0$ (see Theorem 10). □

Remark. For lubrication problems, there is no such singularity on $\partial\Gamma$ because H does not vanish on it (see [2]). One can find, in some oceanic models, the introduction of the same hypothesis $H \geq H_0 > 0$ (see [10, pp. 17–18] and [11, p. 1016]). □

Remark. For Navier–Stokes equations, it is not necessary to give a condition on the shore ∂S since $V^\varepsilon = 0$ on the bottom F ensures that $V^\varepsilon = 0$ on ∂F. Indeed V^ε cannot possess a singularity on ∂S since $(V^\varepsilon, W^\varepsilon) \in (H^1(D))^3$. This is why (9) closes (8). On the contrary the vertical diffusion model (10) allows some singularities on ∂S as it is checked in the following 2D example. □

Example of singularity. Let us consider an infinite channel with a crossing wind: $\Gamma = (0,1) \times \mathbb{R}$, $H = H(X_1)$, $G = (1,0)$. Looking for a solution $\mathbf{V} = (\mathbf{V}_1(X_1), 0)$ of (10) we find, by (13), $\partial_1\mathbf{P} = \frac{3}{2}(1/H + c/H^3)$, where c is an arbitrary real number and, by (11),

$$\mathbf{V}_1 = \frac{1}{4}(3Z^2 + 4ZH + H^2)\frac{1}{H} + \frac{3}{4}(Z^2 - H^2)\frac{c}{H^3}.$$

The flux on each vertical is $\overline{\mathbf{V}}_1 = \int_{-H}^{0} \mathbf{V}_1\, dZ = c/2$, which is constant. If $c \geq 0$, the fluid moves from left to right; therefore, it goes in through the left shore and out through the right shore. This corresponds to a singularity on ∂S, excepted for $c = 0$, which is the only good solution satisfying the "lost condition" (14).

Nevertheless, even for $c \neq 0$, $\partial_Z\mathbf{V}_1$ lies in $L^2(D_1)$, where D_1 is the 2D section of the channel, when $H \geq |X_1(1 - X_1)|^\alpha$ for some $\alpha > 1/3$. □

2. Functional spaces.

2.1. The space $\mathbb{E}$. Throughout the paper, we suppose that

$$\Gamma \text{ is an open bounded subset of } \mathbb{R}^2, \tag{16}$$

$$h \in \mathcal{C}(\overline{\Gamma}),\ h > 0 \text{ in } \Gamma,\ h = 0 \text{ on } \partial\Gamma. \tag{17}$$

We denote

$$\begin{aligned} \mathcal{E}(d) = \{(v,w) \in (\mathcal{C}^\infty(\overline{d}))^2 \times \mathcal{C}^\infty(\overline{d}) :\ & \nabla \cdot v + \partial_z w = 0, \\ & (v,w) = 0 \text{ on a neighborhood of } f,\ w = 0 \text{ on } s\}. \end{aligned} \tag{18}$$

By definition,

$$\begin{cases} \mathbb{E}(d) \text{ is the closure of } \mathcal{E}(d) \text{ for the norm} \\ \|(v,w)\|_{\mathbb{E}(d)} = \displaystyle\left(\int_d |\nabla v|^2 + |\nabla w|^2 + |\partial_z v|^2 + |\partial_z w|^2 \, dxdz\right)^{1/2}. \end{cases}$$

Remark. In the definition (18) we suppose $w = 0$ only at the surface s and not in a neighborhood of s. This is a crucial point in the proof of Lemma 3.

If we suppose in (18) that $w = 0$ in a neighborhood of s, we should suppose then that the force due to the wind satisfies $\nabla \cdot g = 0$. Indeed in the neighborhood we would have $\nabla \cdot v = \nabla \cdot v + \partial_z w = 0$ and thus $\nabla \cdot \partial_z v = 0$; on the other hand, $\nu\, \partial_z v = g$ on s by (2), leading to the conclusion. ∎

LEMMA 1. *The embedding* $\mathbb{E}(d) \subset (H^1(d) \cap L^4(d))^3$ *holds and for any* $(v,w) \in \mathbb{E}(d)$,

$$\|(v,w)\|_{(L^2(d))^3} \le 2^{-1/2} h_{\max} \|\partial_z (v,w)\|_{(L^2(d))^3}, \tag{19}$$

$$\|(v,w)\|_{(L^4(d))^3} \le \gamma_d \|(\nabla, \partial_z)(v,w)\|_{(L^2(d))^9}, \tag{20}$$

where $\gamma_d = 2^{1/8} \beta^{3/4} (h_{\max})^{1/4}$, β *being defined by* (23). *It satisfies* $\beta \le 4$.

Moreover, (v,w) *has a trace in* $(H^{1/2}_{\mathrm{loc}}(s))^3$ *which belongs to* $(L^2(s))^3$ *and satisfies* $w_{|_s} = 0$, *and*

$$\int_s \frac{|v|^2}{h} dx \le \int_d |\partial_z v|^2 \, dxdz. \tag{21}$$

We denote $h_{\max} = \max\{h(x) : x \in \Gamma\}$, and we identify Γ with s to give a meaning to h (or to g) on s.

Remark. By definition, every element of $\mathbb{E}(d)$ is a distribution with derivatives in L^2. This single property does not imply that (v,w) belongs to L^2, or a fortiori to H^1, since d does not possess the cone property. (See the remark following Theorem 4.) ∎

Remark. To be in $\mathbb{E}(d)$ implies, in a weak sense, the condition $(v,w) = 0$ on the bottom f. This bottom is not regular enough (it is only the graph of a continuous function) to define the trace of (v,w) in a usual sense or even to define $L^2(f)$.

On the contrary, the trace on the surface is defined since s is a part of a plane. ∎

Proof of Lemma 1. It suffices to prove the inequalities for $(v,w) \in \mathcal{E}(d)$.

Inequality (19). We have $(v,w)(x,z) = \int_{-h(x)}^{z} \partial_z(v,w)(x,\zeta)\,d\zeta$; hence

$$|(v,w)(x,z)|^2 \le (z+h(x)) \int_{-h(x)}^{0} |\partial_z(v,w)(x,\zeta)|^2\,d\zeta,$$

and therefore

$$\int_{-h(x)}^{0} |(v,w)(x,z)|^2\,dz \le \frac{1}{2}(h(x))^2 \int_{-h(x)}^{0} |\partial_z(v,w)(x,z)|^2\,dz.$$

We conclude by bounding h and integrating with respect to x.

Inequality (20). With the help of Riesz's inequality, we have

$$\|(v,w)\|_{(L^4(d))^3} \le (\|(v,w)\|_{(L^2(d))^3})^{1/4} (\|(v,w)\|_{(L^6(d))^3})^{3/4}.$$

We conclude by bounding the right-hand side with (19) and

$$\|(v,w)\|_{(L^6(d))^3} \le 2^{1/3}\beta \|(\nabla,\partial_z)(v,w)\|_{(L^2(d))^9}. \tag{22}$$

This last inequality comes from the Sobolev–Gagliardo–Nirenberg inequality on $\mathbb{R}^3$ by extending (v,w) by 0 in the half space $\{(x,z) : z \le 0\}$, then by extending $\widetilde{v}$ by symetrization and $\widetilde{w}$ by antisymetrization on z in the whole space $\mathbb{R}^3$; that is, for $z \ge 0$,

$$\widetilde{v}(x,-z) = \widetilde{v}(x,z), \quad \widetilde{w}(x,-z) = -\widetilde{w}(w,z).$$

The function $(\widetilde{v},\widetilde{w})$ is Lipschitz continuous and has a compact support in $\mathbb{R}^3$, from which (see, for instance, [7, p. 162])

$$\|(\widetilde{v},\widetilde{w})\|_{(L^6(\mathbb{R}^3))^3} \le \beta \|(\nabla,\partial_z)(\widetilde{v},\widetilde{w})\|_{(L^2(\mathbb{R}^3))^9}, \tag{23}$$

which yields (22). The inequality $\beta \le 4$ is given in the note (2) on p. 162 of [7].

Inequality (21). We have $v(x,0) = \int_{-h(x)}^{0} \partial_z v(x,z)\,dz$; therefore

$$|v(x,0)|^2 \le h(x) \int_{-h(x)}^{0} |\partial_z v(x,z)|^2\,dz.$$

We conclude by dividing by $h(x)$ and integrating with respect to x. □

2.2. The space $\mathbb{F}$. By definition,

$$\mathbb{F}(d) \text{ is the closure of } \mathcal{E}(d) \text{ for the norm } \|(v,w)\|_{\mathbb{F}(d)} = \left(\int_d |\partial_z v|^2\,dxdz\right)^{1/2}. \tag{24}$$

LEMMA 2. *The space $\mathbb{F}(d)$ is a Hilbert space and*

$$\mathbb{E}(d) \subset \mathbb{F}(d) \subset (L^2(d))^2 \times H^{-1}(d).$$

For all $(v,w) \in \mathbb{F}(d)$

$$\|v\|_{(L^2(d))^2} \le \frac{1}{\sqrt{2}} h_{\max} \|\partial_z v\|_{(L^2(d))^2}, \tag{25}$$

$$\|w\|_{H^{-1}(d)} \le \frac{1}{2}(h_{\max})^2 \|\partial_z v\|_{(L^2(d))^2}, \tag{26}$$

$$\|\partial_z w\|_{H^{-1}(d)} \le \frac{1}{\sqrt{2}} h_{\max} \|\partial_z v\|_{(L^2(d))^2}, \tag{27}$$

and $v/h \in (L^2(d))^2$ *with*

$$\Big\|\frac{v}{h}\Big\|_{(L^2(d))^2} \leq \frac{1}{\sqrt{2}}\|\partial_z v\|_{(L^2(d))^2}. \tag{28}$$

Moreover, v *has a trace in* $(L^2(s))^2$ *which satisfies*

$$\int_s \frac{|v|^2}{h}\,dx \leq \int_d |\partial_z v|^2\,dxdz, \tag{29}$$

and w *has a trace in* $H^{-1}_{\rm loc}(s)$ *which is null.*

We denote

$$H^{-1}(d) = \left\{\varphi \in \mathcal{D}'(d) : \varphi = \varphi_0 + \frac{\partial \varphi_1}{\partial x_1} + \frac{\partial \varphi_2}{\partial x_2} + \frac{\partial \varphi_3}{\partial z},\ \varphi_j \in L^2(d) \text{ for } j \geq 0\right\},$$

this space being endowed with the norm

$$\|\varphi\|_{H^{-1}(d)} = \inf_{\{\varphi_j\}} \left(\int_d \varphi_0^2 + \cdots + \varphi_3^2\right)^{1/2}, \tag{30}$$

where the infimum is taken over all possible decompositions of φ.

Remark. The space $H^{-1}(d)$ is the dual space of $H^1_0(d)$ and its norm is equivalent to the dual norm, but these norms are not uniformly equivalent when d varies. ∎

Proof of Lemma 2. It suffices to prove the inequalities $\forall (v,w)$ in $\mathcal{E}(d)$.

Inequality (28). We have $v(x,z) = \int_{-h(x)}^{z} \partial_z v(x,\zeta)\,d\zeta$, which yields, as in the proof of (19),

$$\int_{-h(x)}^{0} |v(x,z)|^2\,dz \leq \frac{1}{2}(h(x))^2 \int_{-h(x)}^{0} |\partial_z v(x,z)|^2\,dz.$$

We conclude by dividing by $(h(x))^2$ and integrating with respect to x.

Inequality (25). It follows from (28) since $h(x) \leq h_{\max}$.

Inequality (27). We have $\partial_z w = -\nabla \cdot v$; therefore $\|\partial_z w\|_{H^{-1}(d)} \leq \|v\|_{(L^2(d))^2}$. We conclude with (25).

Inequality (26). Denoting by $\widehat{\ }$ the extension by 0 for $z \leq -h(x)$, we have

$$\begin{aligned} w(x,z) &= -\int_{-h(x)}^{z} (\nabla \cdot v)(x,\zeta)\,d\zeta \\ &= -\int_{-h_{\max}}^{z} (\nabla \cdot \widehat{v})(x,\zeta)\,d\zeta \\ &= (\nabla \cdot y)(x,z), \end{aligned}$$

where $y(x,z) = -\int_{-h_{\max}}^{z} \widehat{v}(x,\zeta)\,d\zeta = -\int_{-h(x)}^{z} v(x,\zeta)\,d\zeta$. We bound, as above,

$$\int_d |y(x,z)|^2\,dxdz \leq \frac{1}{2}(h_{\max})^2 \int_d |v(x,\zeta)|^2\,dxd\zeta;$$

therefore

$$\|w\|_{H^{-1}(d)} \leq \|y\|_{(L^2(d))^2} \leq \frac{1}{\sqrt{2}} h_{\max}\|v\|_{(L^2(d))^2},$$

which, together with (25), gives (26).

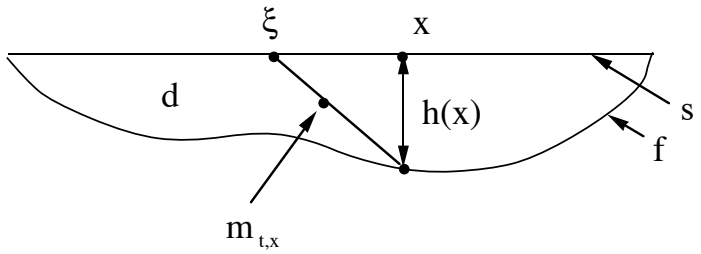


FIG. 2. *A star-shaped domain with respect to its surface.*

Traces. Given $(v,w) \in \mathbb{E}(d)$, $\Gamma'' \subset\subset \Gamma$, and $h'' = \min\{h(x) : x \in \Gamma''\}$, we have $\partial_z w = -\nabla \cdot v \in L^2((-h'',0); H^{-1}(\Gamma''))$. Hence $w \in \mathcal{C}([-h'',0]; H^{-1}(\Gamma''))$, which allows us to define its value for $z = 0$. We obtain a trace in $H^{-1}_{\rm loc}(s)$ by collecting traces in each Γ''.

Likewise, $\partial_z v \in L^2((-h'',0); (L^2(\Gamma''))^2)$ and therefore $v \in \mathcal{C}([-h'',0]; (L^2(\Gamma''))^2)$, which allows us to define the trace in $L^2_{\rm loc}(s)$ by recollection. The inequality (29) follows from (21). It shows that the trace belongs to $L^2(s)$. ∎

2.3. Approximation of $\mathbb{F}$ by smooth functions.

LEMMA 3. *We suppose that*

(31) $$d \text{ is star-shaped with respect to a point of } s.$$

Let $(v,w) \in (\mathcal{D}'(d))^3$ *be such that*

$$\text{(32)} \quad \begin{cases} \exists (\widetilde{v},\widetilde{w}) \in (L^2(\mathbb{R}^3_-))^2 \times \mathcal{D}'(\mathbb{R}^3_-), \quad \partial_z \widetilde{v} \in (L^2(\mathbb{R}^3_-))^2, \quad \nabla \cdot \widetilde{v} + \partial_z \widetilde{w} = 0, \\ (\widetilde{v},\widetilde{w}) = (v,w) \text{ in } d, \quad (\widetilde{v},\widetilde{w}) = 0 \text{ in } \mathbb{R}^3_- \setminus \overline{d}, \quad \widetilde{w} = 0 \text{ on } s. \end{cases}$$

Then $(v,w) \in \mathbb{F}(d)$.

We denote $\mathbb{R}^3_- = \{(x,z) : x \in \mathbb{R}^2, z < 0\}$.

Remark. The hypothesis (31) means (see Figure 2) that there exists $\xi \in \Gamma$ such that

$$m_{t,x} = (tx + (1-t)\xi, -t\,h(x)) \in d \quad \forall x \in \Gamma,\ \forall t \in (0,1).$$

It does not allow us to consider islands, because it implies that Γ is star-shaped with respect to ξ. It no longer allows us to consider two submarine valleys separated by a large plateau close to the surface. ∎

Remark. The trace of $\widetilde{w}$ on s is defined since $\partial_z \widetilde{w} = -\nabla \cdot \widetilde{v}$, which lies in $L^2(\mathbb{R}_-; H^{-1}(\mathbb{R}^2))$. ∎

Remark. The property (32) characterizes $\mathbb{F}(d)$. Indeed it is sufficient by Lemma 3 and it is necessary from the definitions (24) and (18) and from Lemma 2. ∎

Proof of Lemma 3. Let (v,w) satisfy (32). By definition (24) of $\mathbb{F}(d)$, it suffices to find a sequence in $\mathcal{E}(d)$ which is a Cauchy sequence for the norm $\|\partial_z v\|_{(L^2(d))^3}$ and which converges to (v,w) in $(\mathcal{D}'(d))^3$.

The condition (32) gives an extension $(\widetilde{v},\widetilde{w}) \in (L^2(\mathbb{R}^3_-))^2 \times \mathcal{C}(\mathbb{R}_-; H^{-1}(\mathbb{R}^2))$ such that $\widetilde{w} = 0$ for $z = 0$. Thus we define an extension $(\widetilde{v},\widetilde{w}) \in (L^2(\mathbb{R}^3))^2 \times \mathcal{C}(\mathbb{R}; H^{-1}(\mathbb{R}^2))$ in the whole space $\mathbb{R}^3$ by

$$(\widetilde{v},\widetilde{w})(x,z) = \big(\widetilde{v}(x,-z), -\widetilde{w}(x,-z)\big) \quad \text{if } z > 0.$$

By (31), d is star-shaped with respect to a point of s, which is chosen to be the origin. Given $\eta > 0$, we define

$$(v^\eta, w^\eta)(x,z) = (\widetilde{v},\widetilde{w})\big((1+\eta)x,(1+\eta)z\big).$$

Then, $\nabla \cdot v^\eta + \partial_z w^\eta = 0$ and support $(v^\eta, w^\eta) \subset \widetilde{d}$, where $\widetilde{d} = \{(x,z) :\ x \in \Gamma, |z| < h(x)\}$.

Let $\rho^\eta \in \mathcal{D}(\mathbb{R}^3)$ be a mollifier such that

$$\rho^\eta(x,z) = \rho^\eta(x,-z), \quad \int_{\mathbb{R}^3} \rho^\eta \, dxdz = 1, \quad \text{support } \rho^\eta + \text{support}\,(v^\eta, w^\eta) \subset \widetilde{d}.$$

By convolution, we obtain a function $(v^\eta \star \rho^\eta, w^\eta \star \rho^\eta)$ lying in $\mathcal{E}(d)$. Indeed, this function is $\mathcal{C}^\infty$ on $\overline{d}$, it satisfies $\nabla \cdot (v^\eta \star \rho^\eta) + \partial_z (w^\eta \star \rho^\eta) = 0$, it vanishes in a neighborhood of F, and $w^\eta \star \rho^\eta$ vanishes on s because $(w^\eta \star \rho^\eta)(x,0)$ is the integral on $\mathbb{R}^3$ of an antisymetric function with respect to z.

As $\eta \to 0$, we have $\partial_z(v^\eta \star \rho^\eta) \to \partial_z v$ in $(L^2(d))^2$; therefore $(v^\eta \star \rho^\eta, w^\eta \star \rho^\eta) \to (v,w)$ in $\mathbb{F}(d)$ and, by Lemma 2, in $(\mathcal{D}'(d))^3$. ∎

3. Solution of the Navier–Stokes equations. We define here a variational solution of the Navier–Stokes equations (1) and (2), for which we will pass to the limit for $\varepsilon \to 0$ further. Throughout this paper, we suppose that

$$h^{1/2} g \in (L^2(\Gamma))^2. \tag{33}$$

In order to recover a unique pressure, we introduce a nonempty bounded set

$$d' = \Gamma' \times (-h', 0), \quad \Gamma' \subset\subset \Gamma, \quad h' = \min_{x \in \Gamma'} h(x), \tag{34}$$

and we impose

$$\int_{d'} p \, dxdz = 0. \tag{35}$$

Now we are able to state the existence and uniqueness result.

THEOREM 4. *Suppose that* (16), (17), *and* (33) *are satisfied.*

(i) *There exists a solution* $(v,w) \in \mathbb{E}(d)$ *of the following:* $\forall (\varphi,\psi) \in \mathbb{E}(d)$,

$$\left\{ \begin{aligned} &\int_d \nu \left(\nabla v \cdot \nabla \varphi + \nabla w \cdot \nabla \psi + \partial_z v \cdot \partial_z \varphi + \partial_z w \, \partial_z \psi\right) + k'(w\,\varphi_1 - v_1\,\psi) \\ &\qquad + kv \times \varphi + (v \cdot \nabla + w\,\partial_z)\, v \cdot \varphi + (v \cdot \nabla + w\,\partial_z)\, w\,\psi \, dxdz = \int_s g \cdot \varphi \, dx. \end{aligned} \right. \tag{36}$$

(ii) *Every solution satisfies*

$$\int_d |\nabla v|^2 + |\nabla w|^2 + |\partial_z v|^2 + |\partial_z w|^2 \, dxdz \le \frac{1}{\nu^2} \int_\Gamma h\,|g|^2 \, dx. \tag{37}$$

(iii) *To every solution corresponds a unique pressure* $p \in L^2_{\text{loc}}(d)$ *such that the strong equation* (1) *is satisfied in the distribution sense and the condition* (35) *holds.*

It satisfies $\nabla p \in (H^{-1}(d))^2$, $\partial_z p \in H^{-1}(d)$ *and, for every Lipschitz set* d'' *included in* d,

$$\|\nabla p\|_{(H^{-1}(d))^2} + \|\partial_z p\|_{H^{-1}(d)} \le \left(1 + \frac{|\omega|}{\sqrt{2}\nu} h_{\max}\right) \|h^{1/2} g\|_{(L^2(\Gamma))^2} + \left(\frac{\gamma_d}{\nu} \|h^{1/2} g\|_{(L^2(\Gamma))^2}\right)^2$$

$$\|p\|_{L^2(d'')} \le c_{d,d',d''} \left(\left(1 + \frac{|\omega|}{\sqrt{2}\nu} h_{\max}\right) \|h^{1/2} g\|_{(L^2(\Gamma))^2} + \left(\frac{\gamma_d}{\nu} \|h^{1/2} g\|_{(L^2(\Gamma))^2}\right)^2\right).$$

(iv) *There is a unique solution* (v, w, p) *if*

$$\nu^2 > 2^{1/4}\beta^{3/2}(h_{\max})^{1/2}\|h^{1/2}g\|_{(L^2(\Gamma))^2}.$$

(v) *Assume in addition that* h *is Lipschitz continuous on* Γ *and* $g \in (H^1_{\mathrm{loc}}(\Gamma))^2$.

Then the boundary conditions (2) *are satisfied in the trace sense, and* (36) *is equivalent to* (1), (2), *and* $(v, w) \in \mathbb{E}(d)$.

We denote by $\times$ the vector product in $\mathbb{R}^2$; therefore $v \times \varphi = v_1\varphi_2 - v_2\varphi_1 = Bv \cdot \varphi$, and γ_d and β are defined in Lemma 1.

Remark. Without the assumptions of part (v), there is not enough regularity on $\partial_z v$ to define its trace on s and on the bottom f to define a trace on it.

Notice that the domain d does not necessarily possess the cone property, even if h is Lipschitz continuous. Indeed the cone property may fail on the shore ∂S, unless we assume in addition some regularity on $\partial\Gamma$ (the 2D cone property) and that the slope of the bottom does not cancel on the shore (for instance, in any point of $\partial\Gamma$, ∇h is continuous and does not vanish). We do not use such an assumption since, for real lakes, the slope may vanish on the shore and the cone condition may fail. □

Remark. The condition (35) defines a unique p if d, and therefore Γ, is connected. Otherwise, an open set d' must be introduced in each connected component of d, and (35) must be imposed on each d'. □

Remark. The regularity hypothesis (33) on g may be weakened in the interior of Γ: there exists a solution as soon as $g = g' + g''$ with g' satisfying (33) and $g'' \in (H^{-1/2}(\Gamma))^2$ with support included in Γ. Similarly, the results of part (v) probably hold for $g \in (H^{1/2}_{\mathrm{loc}}(\Gamma))^2$. □

Proof of Theorem 4. *Part* i. We will proceed by linearization and fixed point. In order to shorten, we denote

$$u = (v_1, v_2, w), \quad D = (\partial_{x_1}, \partial_{x_2}, \partial_z).$$

Linearization. Given $u' \in \mathbb{E}(d)$, the Lax–Milgram theorem gives the existence of a unique $u \in \mathbb{E}(d)$ such that for any $\mu \in \mathbb{E}(d)$,

$$\int_d \nu Du \cdot D\mu + \omega \times u \cdot \mu + (u' \cdot D)u \cdot \mu \, dxdz = \int_s g \cdot \varphi \, dx, \tag{38}$$

where $\varphi = (\mu_1, \mu_2)$, $\omega = (0, k', k)$, and $\omega \times u = (k'w - kv_2, kv_1, -k'v_1)$.

Indeed, the left-hand side defines a continuous bilinear form with respect to u and μ since, by (20), for any u', u and μ in $\mathbb{E}(d)$,

$$\left|\int_d (u' \cdot D)u \cdot \mu \, dxdz\right| \le (\gamma_d)^2 \|Du'\|_{(L^2(d))^9}\|Du\|_{(L^2(d))^9}\|D\mu\|_{(L^2(d))^9}.$$

This form is coercive since $\omega \times u \cdot u = 0$ and

$$\int_d (u' \cdot D)u \cdot u \, dxdz = 0. \tag{39}$$

To prove this last equation, it suffices to consider u' and u in $\mathcal{E}(d)$. Then, introducing

$d_{h_{\max}} = \Gamma \times (-h_{\max}, 0)$ and denoting by $\widetilde{u}$ the extension by 0 in $d_{h_{\max}} \setminus d$, we get

$$\begin{aligned}\int_d (u' \cdot D)u \cdot u \, dxdz &= \int_{d_{h_{\max}}} (\widetilde{u'} \cdot D)\, \widetilde{u} \cdot \widetilde{u} \, dxdz \\ &= \frac{1}{2} \int_{d_{h_{\max}}} D \cdot (\widetilde{u'}\, |\widetilde{u}|^2)\, dxdz \\ &= \frac{1}{2} \int_s w' \, |u|^2 \, dx \\ &= 0\end{aligned}$$

since $\widetilde{u'} \cdot n_{\partial d_{h_{\max}}} = w' = 0$ on s and $\widetilde{u'} = 0$ on the remainder of $\partial d_{h_{\max}}$. We used $D \cdot \widetilde{u'} = 0$, which follows from the definition of $\mathcal{E}(d)$.

On the other hand, the right-hand side of (38) defines a continuous linear form with respect to μ since, due to (21), we have for any $\mu \in \mathbb{E}(d)$,

$$\begin{aligned}\left| \int_s g \cdot \varphi \, dx \right| &\le \left(\int_s h \, |g|^2 \, dx \right)^{1/2} \left(\int_s \frac{|\varphi|^2}{h} \, dx \right)^{1/2} \\ &\le \|h^{1/2} g\|_{(L^2(\Gamma))^2} \|\partial_z \varphi\|_{(L^2(d))^2}.\end{aligned} \tag{40}$$

The space $\mathbb{E}(d)$ is a Hilbert space: it is a complete space by definition and separated by (20). Therefore the Lax–Milgram theorem ensures the existence and the uniqueness of u, and

$$\|Du\|_{(L^2(d))^9} \le \frac{1}{\nu} \|h^{1/2} g\|_{(L^2(\Gamma))^2}. \tag{41}$$

Fixed point. We denote

$$C = \left\{ u \in \mathbb{E}(d) : \|Du\|_{(L^2(d))^9} \le \frac{1}{\nu} \|h^{1/2} g\|_{(L^2(\Gamma))^2} \right\}.$$

It is a convex, nonempty, and compact set in $(L^4(d))^3$. Moreover, $u' \mapsto u$ maps C into itself by (41), and it is continuous from C (endowed with the norm $(L^4(d))^3$) into itself as we will see. Therefore Schauder's theorem gives the existence of a fixed point. This one satisfies (38) with $u' = u$, that is (36).

It remains to check the continuity. Let $\mathbf{u}'$ be another element of C and $\mathbf{u}$ be the corresponding solution of (38). For all $\mu \in \mathbb{E}(d)$ we have

$$\int_d \nu D(\mathbf{u} - u) \cdot D\mu + \omega \times (\mathbf{u} - u) \cdot \mu + (\mathbf{u}' \cdot D)\, (\mathbf{u} - u) \cdot \mu + \big((\mathbf{u}' - u') \cdot D\big)\, u \cdot \mu \, dxdz = 0.$$

For $\mu = \mathbf{u} - u$ we get, together with (39) and (20),

$$\begin{aligned}\nu \int_d |D(\mathbf{u} - u)|^2 \, dxdz &= - \int_d \big((\mathbf{u}' - u') \cdot D\big)\, u \cdot (\mathbf{u} - u) \, dxdz \\ &\le \gamma_d \, \|\mathbf{u}' - u'\|_{(L^4(d))^3} \|Du\|_{(L^2(d))^9} \|D(\mathbf{u} - u)\|_{(L^2(d))^9}.\end{aligned}$$

With the help of (41) and, once again, (20) we conclude that

$$\|\mathbf{u} - u\|_{(L^4(d))^3} \le \frac{1}{\nu^2} (\gamma_d)^2 \, \|h^{1/2} g\|_{(L^2(\Gamma))^2} \|\mathbf{u}' - u'\|_{(L^4(d))^3}. \tag{42}$$

Part ii. It suffices to choose $(\varphi,\psi)=(v,w)$ in (36) (which is (38) with $u'=u$) and to use (39) and (40) to obtain the announced estimate.

Part iii. Let us choose $(\varphi,\psi)=\mu\in(\mathcal{D}(d))^3$ such that $D\cdot\mu=0$. Then (36) gives

$$\langle -\nu D^2u+\omega\times u+(u\cdot D)u,\mu\rangle_{(\mathcal{D}'(d))^3\times(\mathcal{D}(d))^3}=0;$$

therefore de Rham's theorem (see, for instance, [13]) gives the existence of $p\in\mathcal{D}'(d)$ satisfying $-\nu D^2u+\omega\times u+(u\cdot D)u=-Dp$, that is, (1). This equation can be written, since $D\cdot u=0$, as

$$Dp=-\omega\times u+D\cdot(\nu Du-u\otimes u); \tag{43}$$

therefore, from the definition (30) of the norm in $H^{-1}(d)$ and (19), (20),

$$\begin{aligned}\|Dp\|_{(H^{-1}(d))^3}&\le\|\omega\times u\|_{(L^2(d))^3}+\|\nu Du-u\otimes u\|_{(L^2(d))^9}\\&\le|\omega|\frac{h_{\max}}{\sqrt{2}}\|Du\|_{(L^2(d))^9}+\nu\|Du\|_{(L^2(d))^9}+(\gamma_d\|Du\|_{(L^2(d))^9})^2,\end{aligned}$$

which, with (41), gives the announced estimates on ∇p and $\partial_z p$.

It follows (see, for instance, [14, Theorem 14], with $g=1_{d'}$) that p lies in $L^2(d'')$ for any Lipschitz set d'' included in d, with

$$\|p\|_{L^2(d'')}\le c_{d,d',d''}\left(\|Dp\|_{(H^{-1}(d))^3}+\left|\int_{d'}p\,dxdz\right|\right). \tag{44}$$

Since Dp is unique, p is unique up to a constant which is given by (35), then p is unique.

Part iv. The difference between two solutions u and $\mathbf{u}$ satisfies (42) with $u'=u$ and $\mathbf{u}'=\mathbf{u}$. Thus, it vanishes as soon as $(\gamma_d)^2\|h^{1/2}g\|_{(L^2(\Gamma))^2}<\nu^2$, which gives the announced condition due to the definition of γ_d given in Lemma 1.

Part v. Now $g\in(H^1_{\rm loc}(\Gamma))^2$; thus $u\in(H^2_{\rm loc}(d\cup s))^3$ and $p\in H^1_{\rm loc}(d\cup s)$. This regularity up to the surface may be proved by using either the translation method for the mixed formulation as in [9], the representation by Green's function as in [3], or the Agmon–Douglis–Nirenberg method as in [15, pp. 33–35]. (The reader is referred to these authors since this part of Theorem 4 is not used in the following.) Then, the following Green and Stokes's formulas hold for any $\mu=(\phi,\psi)$ in $\mathcal{E}(d)$, and therefore for any μ in $\mathbb{E}(d)$,

$$\int_d Du\cdot D\mu+D^2u\cdot\mu\,dxdz=\int_s\partial_z v\cdot\phi\,dx,\quad\int_d Dp\cdot\mu\,dxdz=0. \tag{45}$$

Substracting ν-times the first formula from (36) (that is, from (38) with $u'=u$) and adding the second formula, we get

$$\int_d\big(-\nu D^2u+\omega\times u+(u\cdot D)u+Dp\big)\cdot\mu\,dxdz=\int_s(g-\nu\partial_z)\cdot\varphi\,dx. \tag{46}$$

The left-hand side vanishes from (1), and since $\varphi_{|s}$ spans $(\mathcal{D}(s))^2$ this implies that $\nu\partial_z v=g$ on s.

On the other hand, here the bottom f is locally the graph of the Lipschitz function h, and then $(v,w)\in\mathbb{E}(d)$ gives $(v,w)=0$ on f in the trace sense. Finally, $w=0$ on s by Lemma 2; thus all the boundary conditions (2) are proven in the trace sense.

Conversely, if (1) and (2) hold, (46) holds and we get the variational equation (36) by using (45). □

4. Solution of the vertical diffusion model. We prove here the existence and uniqueness of a solution, and we give an equivalent variational formulation. Now we work in the transported domain D (one can choose $D = d_{\rm true}$). The hypotheses (17) and (33) are then equivalent to

$$H \in \mathcal{C}(\overline{\Gamma}),\ H > 0 \text{ in } \Gamma,\ H = 0 \text{ on } \partial\Gamma, \tag{47}$$

$$H^{1/2} G \in (L^2(\Gamma))^2. \tag{48}$$

THEOREM 5. *We suppose* (16), (47), *and* (48).

(i) *There exists a unique solution* $(\mathbf{V}, \mathbf{W}, \mathbf{P})$ *of*

$$\begin{cases} (\mathbf{V}, \mathbf{W}) \in \mathbb{F}(D), \quad \mathbf{P} \in H^1_{\rm loc}(\Gamma), \quad \displaystyle\int_{\Gamma'} \mathbf{P}\, dX = 0, \\ -\partial_Z^2 \mathbf{V} + \nabla \mathbf{P} = 0, \quad \partial_Z \mathbf{P} = 0, \quad \partial_Z \mathbf{V}_{|S} = G, \end{cases} \tag{49}$$

where Γ' *is given by* (34).

(ii) *The solution of* (49) *satisfies the following:* $\forall (\Phi, \Psi) \in \mathbb{F}(D)$,

$$\int_D \partial_Z \mathbf{V} \cdot \partial_Z \Phi \, dX dZ = \int_S G \cdot \Phi \, dX. \tag{50}$$

Conversely, if $(\mathbf{V}, \mathbf{W})$ *is the unique solution of* (50), *there exists a unique* $\mathbf{P}$ *such that* (49) *is satisfied.*

(iii) *The solution of* (49) *satisfies*

$$\int_D |\partial_Z \mathbf{V}|^2 \, dX dZ \le \int_\Gamma H |G|^2 \, dX.$$

The boundary conditions $\mathbf{W} = 0$ on S, $(\mathbf{V}, \mathbf{W}) = 0$ on F, and $(\int_{-H}^0 \mathbf{V}\, dZ) \cdot n_{\partial\Gamma} = 0$ announced in (10) and (14) are included in a weak sense in the condition $(\mathbf{V}, \mathbf{W}) \in \mathbb{F}(D)$. The traction condition has some sense according to the following result, which gives, in addition, a Green formula and the regularity of $\mathbf{P}$.

LEMMA 6. *Let* $\mathbf{V} \in (L^2(D))^2$ *and* $\mathbf{P} \in \mathcal{D}'(D)$ *be such that* $\partial_Z \mathbf{V} \in (L^2(D))^2$, $\partial_Z \mathbf{P} = 0$ *and* $-\partial_Z^2 \mathbf{V} + \nabla \mathbf{P} = 0$.

For all $\Gamma'' \subset\subset \Gamma$ *and for* $H'' = \min\{H(X) : X \in \Gamma''\}$, *such a function satisfies* $\mathbf{V} \in \mathcal{C}^\infty([-H'', 0]; (L^2(\Gamma''))^2)$, *which allows us to define a trace* $\partial_Z \mathbf{V} \in (L^2_{\rm loc}(\Gamma))^2$.

Moreover $\mathbf{P} \in H^1_{\rm loc}(\Gamma)$ *and the following holds:* $\forall\ (\Phi, \Psi) \in \mathcal{E}(D)$,

$$\int_D \partial_Z \mathbf{V} \cdot \partial_Z \Phi + \partial_Z^2 \mathbf{V} \cdot \Phi \, dX dZ = \int_S \partial_Z \mathbf{V} \cdot \Phi \, dX, \tag{51}$$

$$\int_D \nabla \mathbf{P} \cdot \Phi \, dX dZ = 0. \qquad \square \tag{52}$$

Proof of Lemma 6. *Regularity of* $\mathbf{V}$. The equation gives $\partial_Z^n \mathbf{V} = 0$ for any $n \ge 3$, which implies $\mathbf{V} \in C^\infty([-H'', 0]; (L^2(\Gamma''))^2)$ since $\mathbf{V} \in L^2([-H'', 0]; (L^2(\Gamma''))^2)$.

Regularity of $\mathbf{P}$. The equation gives $\nabla \mathbf{P} = \partial_Z^2 \mathbf{V} \in C([-H'', 0]; (L^2(\Gamma''))^2)$. Since $\nabla \mathbf{P}$ is independent on Z, it belongs to $(L^2(\Gamma''))^2$, and therefore $\mathbf{P} \in H^1_{\rm loc}(\Gamma)$.

Green formula. We cover Γ by a collection of cells Γ_i'' small enough to have $\Phi = 0$ if $X \in \Gamma_i''$, $Z \le -H_i''$. We have $\partial_Z \mathbf{V} \cdot \partial_Z \Phi + \partial_Z^2 \mathbf{V} \cdot \Phi = \partial_Z(\partial_Z \mathbf{V} \cdot \Phi)$ in $C([-H_i'', 0]; (L^2(\Gamma_i''))^2)$. Therefore, in $(L^2(\Gamma_i''))^2$,

$$\int_{-H_i''}^0 \partial_Z \mathbf{V} \cdot \partial_Z \Phi + \partial_Z^2 \mathbf{V} \cdot \Phi \, dZ = (\partial_Z \mathbf{V} \cdot \Phi)_{|Z=0}.$$

We deduce (51) by integrating with respect to X over Γ_i'' and summing with respect to i.

Proof of (52). The extension of Φ by 0 satisfies, for any fixed Z, $\Phi(\cdot, Z) \in (\mathcal{D}(\Gamma))^2$; thus

$$\int_\Gamma \nabla \mathbf{P}(X) \cdot \Phi(X,Z)\, dX = -\int_\Gamma \mathbf{P}(X)\, \nabla \cdot \Phi(X,Z)\, dX.$$

By definition of $\mathcal{E}(D)$, $\nabla \cdot \Phi = -\partial_Z \Psi$; therefore

$$\begin{aligned}\int_D \nabla \mathbf{P}(X) \cdot \Phi(X,Z)\, dX dZ &= \int_D \mathbf{P}(X)\, \partial_Z \Psi(X,Z)\, dX dZ \\ &= \int_\Gamma \mathbf{P}(X) \Big(\int_{-H(X)}^0 \partial_Z \Psi(X,Z)\, dZ \Big)\, dX,\end{aligned}$$

which cancels since $\Psi(X,0) = \Psi(X,-H(X))$. □

Proof of Theorem 5. *Variational solution.* The left-hand side of (50) defines a continuous bilinear form on $\mathbb{F}(D)$, which is coercive since $\int_D |\partial_Z \mathbf{V}|^2\, dX dZ = (\|(\mathbf{V},\mathbf{W})\|_{\mathbb{F}(D)})^2$. The right-hand side of (50) defines a continuous linear form according to (40). Then, the Lax–Milgram theorem gives the existence and uniqueness of a solution $(\mathbf{V},\mathbf{W}) \in \mathbb{F}(D)$ of (50).

Strong solution. Let us choose $(\Phi,\Psi) \in (\mathcal{D}(D))^3$ such that $\nabla \cdot \Phi + \partial_Z \Psi = 0$. Equation (50) yields

$$\langle (-\partial_Z^2 \mathbf{V}, 0), (\Phi,\Psi) \rangle_{(\mathcal{D}'(D))^3 \times (\mathcal{D}(D))^3} = 0;$$

therefore de Rham's theorem (see, for instance, [13]) gives the existence of $\mathbf{P} \in \mathcal{D}'(D)$ such that $(\nabla \mathbf{P}, \partial_Z \mathbf{P}) = (\partial_Z^2 \mathbf{V}, 0)$. Then, Lemma 6 gives $\mathbf{P} \in H^1_{\mathrm{loc}}(D)$. Since $\nabla \mathbf{P}$ is unique, $\mathbf{P}$ is unique up to a constant which is given by $\int_{\Gamma'} \mathbf{P}\, dX = 0$, and therefore $\mathbf{P}$ is unique.

Let us choose now $(\Phi,\Psi) \in \mathcal{E}(D)$. Substracting (50) from (51), one obtains

$$\int_S (\partial_Z \mathbf{V} - G) \cdot \Phi\, dX = \int_D \partial_Z^2 \mathbf{V} \cdot \Phi\, dX dZ.$$

The right-hand side equals $\int_D \nabla \mathbf{P} \cdot \Phi\, dX dZ$, which vanishes due to (52). Since $\Phi_{|S}$ spans $(\mathcal{D}(S))^2$, it follows that $\partial_Z \mathbf{V} = G$.

Converse assertion. Let us consider now a solution of (49) and $(\Phi,\Psi) \in \mathcal{E}(D)$. Since $\partial_Z^2 \mathbf{V} - \nabla \mathbf{P} = 0$,

$$\int_D \partial_Z \mathbf{V} \cdot \partial_Z \Phi\, dX dZ = \int_D \partial_Z \mathbf{V} \cdot \partial_Z \Phi + \partial_Z^2 \mathbf{V} \cdot \Phi - \nabla \mathbf{P} \cdot \Phi\, dX dZ.$$

According to (51) and (52), the right-hand side is equal to $\int_S \partial_Z \mathbf{V} \cdot \Phi\, dX$ and therefore, with the traction condition, to $\int_\Gamma G \cdot \Phi\, dX$ which gives (50). By continuity, this equation is satisfied for all (Φ,Ψ) in $\mathbb{F}(D)$. □

5. Convergence as $\varepsilon \to 0$. We prove here that the solution of the transported Navier–Stokes equations on the fixed domain D goes to the solution of the vertical diffusion model.

THEOREM 7. *Let* (3), (16), (47), *and* (48) *be satisfied.*

As $\varepsilon \to 0$, the image $(V^\varepsilon, W^\varepsilon, P^\varepsilon)$ by the scaling (6) *and* (7) *of the solution $(v^\varepsilon, w^\varepsilon, p^\varepsilon)$ defined by Theorem* 4 *goes to the solution $(\mathbf{V}, \mathbf{W}, \mathbf{P})$ of* (49) *in the following sense*:

$$(V^\varepsilon, W^\varepsilon, P^\varepsilon) \rightharpoonup (\mathbf{V}, \mathbf{W}, \mathbf{P}) \quad \text{in } (L^2(D))^2 \times H^{-1}(D) \times L^2_{\mathrm{loc}}(D) \text{ weak.}$$

Remark. From Theorem 4, $(v^\varepsilon, w^\varepsilon, p^\varepsilon)$, and therefore $(V^\varepsilon, W^\varepsilon, P^\varepsilon)$ is unique as soon as $\varepsilon < \nu^2/\big(2^{1/4}\beta^{3/2}(H_{\max})^{1/2}\|H^{1/2}G\|_{(L^2(\Gamma))^2}\big)$. ∎

Remark. For the limit problem, there is a "strong" formulation, namely (49), which is not the case for the initial problem (under the same hypothesis).

As we mentioned in the remark following Theorem 4, for the Navier–Stokes equations there is not enough regularity to define the trace of $\partial_Z V^\varepsilon$ on S. On the contrary, for the vertical diffusion model, this trace is defined at Lemma 6 with the help of the hydrostatic balance (that is, $\mathbf{P}$ independent on Z) which implies that $\partial_Z^2 \mathbf{V}$ is integrable with respect to Z. ∎

We will deduce Theorem 7 from the following properties, which are easy consequences of the results for a fixed ε.

PROPOSITION 8. *The image by the scaling* (3), (6), *and* (7) *of every solution $(v^\varepsilon, w^\varepsilon, p^\varepsilon)$ defined at Theorem* 4 *satisfies $(V^\varepsilon, W^\varepsilon) \in \mathbb{E}(D)$ and $\forall (\Phi, \Psi) \in \mathbb{E}(D)$,*

$$\left\{ \begin{aligned} &\int_D \partial_Z V^\varepsilon \cdot \partial_Z \Phi + \varepsilon^2 \nabla V^\varepsilon \cdot \nabla \Phi + \varepsilon^2 \partial_Z W^\varepsilon \partial_Z \Psi + \varepsilon^4 \nabla W^\varepsilon \cdot \nabla \Psi \\ &\quad + \frac{\varepsilon^3}{\nu^2}(V^\varepsilon \cdot \nabla + W^\varepsilon \partial_Z) V^\varepsilon \cdot \Phi + \frac{\varepsilon^5}{\nu^2}(V^\varepsilon \cdot \nabla + W^\varepsilon \partial_Z) W^\varepsilon \Psi \, dX dZ \\ &\quad + \frac{\varepsilon^2 k}{\nu} V^\varepsilon \times \Phi + \frac{\varepsilon^3 k'}{\nu}(W^\varepsilon \Phi_1 - V_1^\varepsilon \Psi) = \int_\Gamma G \cdot \Phi \, dX, \end{aligned} \right. \tag{53}$$

$$\int_D |\partial_Z V^\varepsilon|^2 + \varepsilon^2 |\nabla V^\varepsilon|^2 + \varepsilon^2 |\partial_Z W^\varepsilon|^2 + \varepsilon^4 |\nabla W^\varepsilon|^2 \, dX dZ \le \int_\Gamma H |G|^2 \, dX. \tag{54}$$

Moreover, it satisfies the strong equation (8), $P^\varepsilon \in L^2_{\mathrm{loc}}(D)$, *and, for* $\varepsilon \le 1$,

$$\|\nabla P^\varepsilon\|_{(H^{-1}(D))^2} \le \left(1 + \frac{\varepsilon^2 |\omega|}{\sqrt{2}\nu} H_{\max}\right) \|H^{1/2} G\|_{(L^2(\Gamma))^2} + \left(\frac{\gamma_D}{\nu} \|H^{1/2} G\|_{(L^2(\Gamma))^2}\right)^2 \tag{55}$$

$$\|\partial_Z P^\varepsilon\|_{H^{-1}(D)} \le \varepsilon \left(\left(1 + \frac{\varepsilon^2 |\omega|}{\sqrt{2}\nu} H_{\max}\right) \|H^{1/2} G\|_{(L^2(\Gamma))^2} + \left(\frac{\gamma_D}{\nu} \|H^{1/2} G\|_{(L^2(\Gamma))^2}\right)^2 \right) \tag{56}$$

$$\int_{D'} P^\varepsilon \, dX dZ = 0. \tag{57}$$

Here $D' = \Gamma' \times (-H', 0)$, Γ' is given by (34) *and $H' = \min\{H(X) : X \in \Gamma'\}$.*

Proof of Proposition 8. Equation (53) follows from (36) by observing that, if $(\varphi, \psi) \in \mathbb{E}(d_\varepsilon)$, an element $(\Phi, \Psi) \in \mathbb{E}(D)$ is defined by $\Phi(X, Z) = \varphi(x, z)$, $\Psi(X, Z) = \varepsilon \, \psi(x, z)$.

Inequality (54) follows from (37).

To get (55) and (56), we observe that, once scaled by (7), (43) may be written in the following form:

$$\left(\nabla, \frac{1}{\varepsilon}\partial_Z\right) P^\varepsilon = \varepsilon^2 \frac{\omega}{\nu} \times (V^\varepsilon, \varepsilon W^\varepsilon) + (\nabla, \partial_Z) \cdot \Big((\varepsilon^2 \nabla, \partial_Z)(V^\varepsilon, \varepsilon W^\varepsilon) - \frac{\varepsilon^3}{\nu^2} (V^\varepsilon, W^\varepsilon) \otimes (V^\varepsilon, \varepsilon W^\varepsilon) \Big);$$

therefore, with (30), (19), (20), and $\varepsilon \le 1$,

$$\left\| \left(\nabla, \frac{1}{\varepsilon}\partial_Z \right) P^\varepsilon \right\|_{(H^{-1}(D))^3} \le \varepsilon^2 \frac{|\omega|}{\nu} \|(V^\varepsilon, \varepsilon W^\varepsilon)\|_{(L^2(D))^3} + \|(\varepsilon^2\nabla, \partial_Z)(V^\varepsilon, \varepsilon W^\varepsilon)\|_{(L^2(D))^9} + \left(\frac{\varepsilon}{\nu} \|(V^\varepsilon, \varepsilon W^\varepsilon)\|_{(L^4(D))^9} \right)^2$$

$$\le \left(\frac{\varepsilon^2|\omega|}{\sqrt{2}\nu} H_{\max} + 1 \right) \|(\varepsilon\nabla, \partial_Z)(V^\varepsilon, \varepsilon W^\varepsilon)\|_{(L^2(D))^9} + \left(\frac{\gamma_D}{\nu} \|(\varepsilon\nabla, \partial_Z)(V^\varepsilon, \varepsilon W^\varepsilon)\|_{(L^2(D))^9} \right)^2,$$

which, due to (54), gives (55) and (56). ∎

Proof of Theorem 7. *Convergence of the velocity.* From (54), $(V^\varepsilon, W^\varepsilon)$ is bounded in $\mathbb{F}(D)$; therefore there exists a subsequence and $(\mathbf{V}, \mathbf{W})$ such that

$$(V^\varepsilon, W^\varepsilon) \rightharpoonup (\mathbf{V}, \mathbf{W}) \quad \text{in } \mathbb{F}(D) \text{ weak.}$$

Let us pass to the limit in (53) for $(\Phi, \Psi) \in \mathcal{E}(D)$. The weak convergence gives

$$\int_D \partial_Z V^\varepsilon \cdot \partial_Z \Phi \, dX dZ \to \int_D \partial_Z \mathbf{V} \cdot \partial_Z \Phi \, dX dZ.$$

The other terms in the left-hand side of (53) go to 0. Indeed, from (54), $\partial_Z V^\varepsilon$, $\varepsilon\nabla V^\varepsilon$, $\varepsilon\partial_Z W^\varepsilon$, and $\varepsilon^2\nabla W^\varepsilon$ are bounded in $(L^2(D))^n$ ($n = 1$, 2, or 4); thus V^ε and $\varepsilon W^\varepsilon$ are bounded in $(L^2(D))^n$ because of (19).

Therefore the limit $(\mathbf{V}, \mathbf{W})$ satisfies the variational equation (50) for any $(\Phi, \Psi) \in \mathcal{E}(D)$. By continuity, it satisfies this equation for any $(\Phi, \Psi) \in \mathbb{F}(D)$; thus, due to Theorem 5, it is unique. Therefore the whole sequence converges.

Convergence of the pressure. Its gradient $(\nabla P^\varepsilon, \partial_Z P^\varepsilon)$ is bounded in $(H^{-1}(D))^3$ because of (55) and (56). With (57) and (44), it results that P^ε is bounded in $L^2_{\text{loc}}(D)$. Then, there exists a subsequence and $\mathbf{P}$ such that

$$P^\varepsilon \rightharpoonup \mathbf{P} \quad \text{in } L^2_{\text{loc}}(D) \text{ weak.}$$

One can pass to the limit in the first equation of (8), the nonlinear term converging to 0 in $(L^1(D))^2$ since V^ε, $\varepsilon\nabla V^\varepsilon$, $\varepsilon W^\varepsilon$, and $\partial_Z V^\varepsilon$ are bounded in $(L^2(D))^n$ for some n. At the limit, $-\partial_Z^2 \mathbf{V} + \nabla \mathbf{P} = 0$.

From (56), $\partial_Z P^\varepsilon \to 0$; thus $\partial_Z \mathbf{P} = 0$. Finally, passing to the limit in (57), we obtain $\int_{D'} \mathbf{P} \, dX dZ = 0$. Due to Lemma 6, $\mathbf{P}$ satisfies (49), and due to Theorem 5 it is unique. Therefore the whole sequence converges, ending the proof of Theorem 7. ∎

6. Semiexplicit solution of the vertical diffusion model. Let us first give a variational solution $\mathbf{P} = \mathbf{P}(X)$ of the pressure equation. We denote

$$\mathbb{A}(\Gamma) = \left\{ \mathbf{P} \in L^2_{\text{loc}}(\Gamma) : H^{3/2}\nabla \mathbf{P} \in (L^2(\Gamma))^2, \int_{\Gamma'} \mathbf{P} \, dX = 0 \right\} \tag{58}$$

(recall that Γ' is not empty and $\Gamma' \subset\subset \Gamma$), a space which is endowed with the norm

$$\|\mathbf{P}\|_{\mathbb{A}(\Gamma)} = \left(\int_\Gamma H^3 |\nabla \mathbf{P}|^2 \, dX \right)^{1/2}.$$

Notice that $\mathbb{A}(\Gamma) \subset H^1_{\rm loc}(\Gamma)$.

PROPOSITION 9. *Let* (16), (47), *and* (48) *be satisfied. There exists a unique solution* $\mathbf{P} \in \mathbb{A}(\Gamma)$ *of the following*: $\forall\, \Xi \in \mathbb{A}(\Gamma)$,

$$\frac{1}{3}\int_\Gamma H^3 \nabla \mathbf{P} \cdot \nabla \Xi \, dX = \frac{1}{2}\int_\Gamma H^2 G \cdot \nabla \Xi \, dX. \tag{59}$$

It satisfies the strong equation (13) *and*

$$\frac{1}{9}\int_\Gamma H^3 |\nabla \mathbf{P}|^2 \, dX \le \frac{1}{4}\int_\Gamma H |G|^2 \, dX. \tag{60}$$

If Γ *is a Lipschitz set*, $\mathbf{P}$ *satisfies the boundary condition* (15) *in* $H^{-1/2}(\partial\Gamma)$.

The velocity is given explicitly in terms of Z and $\nabla \mathbf{P}(X)$ as follows.

THEOREM 10. *Let* (16), (31), (47), *and* (48) *be satisfied. Then*, $\mathbf{P}$ *being defined by* (59), *formulas* (11) *and* (12) *define the unique velocity* $(\mathbf{V}, \mathbf{W})$ *satisfying* (49).

Remark. The velocity is infinitely regular in the vertical direction. Indeed (49) gives $\partial_Z^2 \mathbf{V} = \nabla \mathbf{P}$, $\partial_Z^n \mathbf{V} = 0$ as soon as $n \ge 3$, and $\partial_Z \mathbf{W} = -\nabla \cdot \mathbf{V}$ gives $\partial_Z^n \mathbf{W} = 0$ as soon as $n \ge 4$. We recover this regularity in the explicit expressions (11) and (12). ∎

Remark. The hypothesis (31) is satisfied by d if and only if it is satisfied by D. ∎

Proof of Proposition 9. *Properties of the space* $\mathbb{A}(\Gamma)$. It is a separated space since, according to (44), for any $\Gamma'' \subset\subset \Gamma$, there exists $c_{\Gamma''}$ such that $\forall\, \mathbf{P} \in \mathbb{A}(\Gamma)$,

$$\left(\int_{\Gamma''} \mathbf{P}^2 \, dX\right)^{1/2} \le c_{\Gamma''} \|\mathbf{P}\|_{\mathbb{A}(\Gamma)}.$$

This inequality implies the completion; thus $\mathbb{A}(\Gamma)$ is a Hilbert space.

Solution of the variational equation (59). Its right-hand side is bounded by

$$\left|\int_\Gamma \frac{1}{2} H^2 G \cdot \nabla \Xi \, dX\right| \le \frac{1}{2}\left(\int_\Gamma H |G|^2 \, dX\right)^{1/2} \left(\int_\Gamma H^3 \, |\nabla \Xi|^2 \, dX\right)^{1/2};$$

therefore the Lax–Milgram theorem gives the existence of a unique solution $\mathbf{P} \in \mathbb{A}(\Gamma)$ of (59) and the inequality (60).

Verification of the strong equation (13). Let $\Xi \in \mathcal{D}(\Gamma)$. There exists a (unique) real number c such that $\Xi + c \in \mathbb{A}(\Gamma)$. Since $\nabla(\Xi + c) = \nabla \Xi$, (59) is satisfied by Ξ. This proves (13) in $\mathcal{D}'(\Gamma)$.

Verification of the boundary condition (15). The function $\overline{\mathbf{V}} = -\frac{1}{3} H^3 \nabla \mathbf{P} + \frac{1}{2} H^2 G$ belongs to $(L^2(\Gamma))^2$. It satisfies $\nabla \cdot \overline{\mathbf{V}} = 0$ because of (13). If Γ is a Lipschitz set, we can define $\overline{\mathbf{V}} \cdot n_{\partial\Gamma} \in H^{-1/2}(\partial\Gamma)$ by the following: $\forall \Xi \in H^1(\Gamma)$,

$$\langle \overline{\mathbf{V}} \cdot n_{\partial\Gamma}, \Xi \rangle_{H^{-1/2}(\partial\Gamma) \times H^{1/2}(\partial\Gamma)} = \int_\Gamma \overline{\mathbf{V}} \cdot \nabla \Xi \, dX.$$

The right-hand side is null because of (59) (it suffices to choose c such that $\Xi + c \in \mathbb{A}(\Gamma)$), which gives (15) in $H^{-1/2}(\Gamma)$. ∎

Proof of Theorem 10. Let $\mathbf{P}$ be defined by (59) and $(\mathbf{V}, \mathbf{W})$ be defined by (11) and (12).

Verification that $(\mathbf{V}, \mathbf{W}) \in \mathbb{F}(D)$. By Lemma 3, it suffices to find an extension $(\widetilde{\mathbf{V}}, \widetilde{\mathbf{W}})$ satisfying (32). Let us define

$$\begin{cases} \widetilde{\mathbf{V}} = \widehat{\mathbf{V}} = \frac{1}{2}\widehat{(Z^2 - H^2)\nabla\mathbf{P}} + \widehat{(Z+H)G}, \\ \widetilde{\mathbf{W}} = \nabla\cdot\left(\widehat{(\frac{1}{2}ZH^2 - \frac{1}{6}Z^3)\nabla\mathbf{P}} - \widehat{(\frac{1}{2}Z^2 + HZ)G}\right), \end{cases}$$

where $\widehat{\ }$ (and $\widehat{}$) denote the extension of functions by 0 outside D. Since $H^{3/2}\nabla\mathbf{P}$ and $H^{1/2}G$ lie in $(L^2(\Gamma))^2$ and $|Z| \le |H|$ in D, each of the extended functions lies in $(L^2(D))^2$ and therefore has an extension in $(L^2(\mathbb{R}^3_-))^2$.

The function $\widetilde{\mathbf{V}}$ coincides with $\mathbf{V}$ in D and it vanishes outside D. The distribution $\widetilde{\mathbf{W}}$ coincides with $\mathbf{W}$ in D and it vanishes outside D.

Let us verify that $\partial_Z\widetilde{\mathbf{V}} \in (L^2(\mathbb{R}^3_-))^2$. In the set $D_\infty = \Gamma \times (-\infty, 0)$, one has $\widetilde{\mathbf{V}} = \frac{1}{2}(\widetilde{Z}^2 - H^2)\nabla\mathbf{P} + (\widetilde{Z} + H)G$, where $\widetilde{Z} = \max\{Z, -H(X)\}$. The map $Z \mapsto \widetilde{Z}$ is differentiable into $L^q(D_\infty)$, $\forall q < \infty$, with $\partial_Z\widetilde{Z} = 1_D$. Since $\nabla\mathbf{P}$ and G belong to $(L^2_{\mathrm{loc}}(\Gamma))^2$, $Z \mapsto \widetilde{\mathbf{V}}$ is differentiable into $L^1_{\mathrm{loc}}(D_\infty)$ and

$$\partial_Z\widetilde{\mathbf{V}} = (Z\nabla\mathbf{P} + G)\,1_D.$$

Therefore

$$\begin{aligned} \left|\int_{D_\infty} |\partial_Z\widetilde{\mathbf{V}}|^2\,dX\,dZ\right| &\le \int_\Gamma dX \int_{-H(X)}^0 |Z\nabla\mathbf{P} + G|^2\,dZ \\ &\le \int_\Gamma \frac{2H^3}{3}|\nabla\mathbf{P}|^2 + 2H|G|^2\,dX, \end{aligned}$$

which is finite because of (60); thus $\partial_Z\widetilde{\mathbf{V}} \in (L^2(D_\infty))^2$. In $\mathbb{R}^3_- \setminus \overline{D}$, we have $\partial_Z\widetilde{\mathbf{V}} = 0$; therefore, by recollection of distributions ($\mathbb{R}^3_-$ is the union of the open sets D_∞ and $R^3_- \setminus \overline{D}$)

$$\partial_Z\widetilde{\mathbf{V}} \in (L^2(\mathbb{R}^3_-))^2.$$

On the other hand, in D_∞,

$$\begin{aligned} \partial_Z\widetilde{\mathbf{W}} &= \nabla\cdot\partial_Z\left(\left(\frac{1}{2}\widetilde{Z}H^2 - \frac{1}{6}\widetilde{Z}^3\right)\nabla\mathbf{P} - \left(\frac{1}{2}\widetilde{Z}^2 + H\widetilde{Z}\right)G\right) \\ &= -\nabla\cdot\left(\frac{1}{2}(\widetilde{Z}^2 - H^2)1_D\nabla\mathbf{P} + (\widetilde{Z} + H)1_D G\right) \\ &= -\nabla\cdot\widetilde{\mathbf{V}}. \end{aligned}$$

This equality is also true in $\mathbb{R}^3_- \setminus \overline{D}$ and therefore, by recollection, in the whole set $\mathbb{R}^3_-$.

Let us verify now the condition at the surface on $\widetilde{\mathbf{W}}$. Let $\Gamma'' \subset\subset \Gamma$ and $H'' = \inf\{H(X) : X \in \Gamma''\}$. In $\Gamma'' \times (-H'', 0)$ we have $\widetilde{Z} = Z$; therefore $\widetilde{\mathbf{W}} = \nabla\cdot Y$, where $Y(Z) = (\frac{1}{2}ZH^2 - \frac{1}{6}Z^3)\nabla\mathbf{P} - (\frac{1}{2}Z^2 + HZ)G$. Since $Y \in \mathcal{C}([-H'', 0]; (L^2(\Gamma''))^2)$ and $Y(0) = 0$, we have $\widetilde{\mathbf{W}} \in \mathcal{C}([-H'', 0]; H^{-1}(\Gamma''))$ and $\widetilde{\mathbf{W}} = 0$ on S. This ends the proof of (32).

Verification of the other properties. Differentiating (11) in the distribution sense, we obtain $\partial_Z^2 \mathbf{V} = \nabla \mathbf{P}$. On the other hand, by definition of $\mathbf{P}$, $\partial_Z \mathbf{P} = 0$ and $\int_{\Gamma'} \mathbf{P}\, dX = 0$. Finally, $\partial_Z \mathbf{V} = Z \nabla \mathbf{P} + G$, which is equal to G on S. Thus, (49) is proved. ∎

Remark. Any element of $L^2(D)$ has a unique extension in $L^2(\mathbb{R}^3_-)$ which vanishes outside $\overline{D}$. In $H^{-1}(\mathbb{R}^3_-)$ there is not uniqueness, since it may exist a "mass" on ∂D. This is the reason why we do not denote in the same manner the two extensions by 0 which are $\widehat{\ }$ (the unique extension in $L^2(\mathbb{R}^3_-)$) and $\widetilde{\ }$ (one of the possible extensions in $H^{-1}(\mathbb{R}^3_-)$). ∎

REFERENCES

[1] A. Assemien, G. Bayada, and M. Chambat, *Inertial effects in the asymptotic behaviour of a thin film flow*, Asymptotic Anal., 9 (1994), pp. 177–208.

[2] G. Bayada and M. Chambat, *The transition between the Stokes equations and the Reynolds equation: A mathematical proof*, Appl. Math. Optim., 14 (1986), pp. 73–93.

[3] J. Bello, *L^r regularity for the Stokes and Navier-Stokes problems*, Ann. Mat. Pura Appl., 4 (1996), pp. 187–206.

[4] O. Besson and M. R. Laydi, *Some estimates for the anisotropic Navier–Stokes equations and for the hydrostatic approximation*, RAIRO Modél. Math. Anal. Numér., 26 (1992), pp. 855–865.

[5] D. Bresch, R. Echevarría, J. Lemoine, and J. Simon, *Numerical simulation of shallow lakes: First results*, in NSE-6 Conference, Aman et al., eds., VSP/TEV, Vilnius, 1998.

[6] D. Bresch, J. Lemoine, and J. Simon, *Ecoulement engendré par le vent et la force de Coriolis dans un domaine mince:* I *Cas stationnaire*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 807–812.

[7] H. Brézis, *Analyse fonctionnelle, théorie et applications*, Masson, Paris, 1983.

[8] T. Colin and P. Fabrie, *Rotating fluid at high Rossby number driven by a surface stress: Existence and convergence*, Adv. Differential Equations, 2 (1997), pp. 715–751.

[9] J.-M. Ghidaglia, *Régularité des solutions de certains problèmes aux limites linéaires liés aux équations d'Euler*, Comm. Partial Differential Equations, 9 (1984), pp. 1265–1298.

[10] J.-L. Lions, R. Temam, and S. Wang, *Models for the coupled atmosphere and ocean*, Comput. Mech. Adv., 1 (1993), pp. 5–54.

[11] J.-L. Lions, R. Temam, and S. Wang, *On the equations of the large-scale ocean*, Nonlinearity, 5 (1992), pp. 1007–1053.

[12] G. Raugel and G. Sell, *Navier-Stokes equations in thin 3D domains*: I *Global attractors and global regularity of solutions*, J. Amer. Math. Soc., 6 (1993), pp. 503–568.

[13] J. Simon, *Démonstration constructive d'un théorème de G. de Rham*, C. R. Acad. Sci. Paris Sér. I Math., 316 (1993), pp. 1167–1172.

[14] J. Simon, *Representation of distributions and explicit antiderivatives up to the boundary*, in Progress in Partial Differential Equations: The Metz Surveys, M. Chipot, ed., Longman, London, 1993, pp. 101–116.

[15] R. Temam, *Navier–Stokes Equations. Theory and Numerical Analysis*, 2nd ed., North-Holland, Amsterdam, 1979.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 623–644

SELF-SIMILAR BLOW-UP AND HAUSDORFF DIMENSION ESTIMATES FOR A CLASS OF PARABOLIC FREE BOUNDARY PROBLEMS*

G. S. WEISS†

Abstract. We consider variational solutions of the quenching problem $\partial_t u - \Delta u = -u^\gamma \chi_{\{u>0\}}$ with exponent $\gamma \in (-\frac{1}{3}, 0)$ and solutions of the heat equation with Bernoulli-type condition on the free boundary which arises in combustion theory, $\partial_t v - \Delta v = 0$ in $\{v > 0\}$, $|\nabla v| = 1$ on $\partial\{v > 0\}$. We show that blow-up limits of u and v are backward self-similar solutions and use this to determine the Hausdorff dimension of the free boundaries $\partial\{u > 0\}$ and $\partial\{v > 0\}$.

Key words. free boundary, monotonicity formula, Hausdorff dimension estimate, characterization of blow-up limits, Bernoulli-type free boundary condition, quenching, combustion

AMS subject classifications. 35R35, 35B05, 35B40, 35K55

PII. S0036141097327409

1. Introduction.

In this paper we study the gradient flow in $L^2(\mathbf{R}^n)$ with respect to the energy

$$w \mapsto F(w) := \int_{\mathbf{R}^n} \left(|\nabla w|^2 \ + \ \lambda_+ \, \chi_{\{w>0\}} \, w^p \ + \ \lambda_- \, \chi_{\{w<0\}} \, (-w)^p \right), \tag{1.1}$$

where the exponent $p \in [0, 2)$, λ_+ and λ_- are real parameters, and χ_A denotes the characteristic function of the set A .

Depending on the choice of p, λ_+, and λ_- as well as initial data a solution u may exist or not. For certain sets of parameters the question of global existence, even in a weak sense, has not yet been answered, and for a large class of parameters uniqueness or even an ordering of solutions is unknown. Given a global weak solution in a sense to be specified, our objective is the study of sets of special relevance for this solution: for example, the "free boundary" $\partial\{u > 0\} \cup \partial\{u < 0\}$, the set of singular free boundary points, and the set of horizontal free boundary points in which the behavior in time is dominant. We are interested in the size in terms of the Hausdorff dimension of these sets as well as in regularity properties of the free boundary. Since differentiability of the solution, even directional differentiability, is not known in advance, *blow-up sequences*

$$u_k(t, x) := \frac{u(T + {\rho_k}^2 t, x_0 + \rho_k x)}{{\rho_k}^{\frac{2}{2-p}}},$$

defined with respect to a given point (T, x_0) and with respect to a given sequence $\rho_k \to 0$, prove to be very useful. The underlying reason for this is that blow-up sequences introduce the possibility of different subsequences converging to different "blow-up limits" and thereby allow—usually by indirect arguments—to obtain information on the solution's behavior at a free boundary point *without any knowledge of the free*

*Received by the editors September 15, 1997; accepted for publication July 23, 1998; published electronically April 7, 1999.
http://www.siam.org/journals/sima/30-3/32740.html
†Tokyo Institute of Technology, O-okayama 2-12-1, Meguro-ku, Tokyo, 152 Japan (gw@math.titech.ac.jp).

boundary's direction. Once the blow-up limits have been characterized, the size of the special sets mentioned above can usually be estimated by a standard dimension reduction procedure.

In section 3 we introduce a variational formulation for the parabolic equation associated with the energy (1.1) which involves the time-space first variation with respect to domain variations of the time-integrated energy. This formulation makes it possible to deal in an elegant way with the singular cases $p \in [0,1)$. Another advantage is that the class of nondegenerate variational solutions is closed with respect to convergence in $H^{1,2}$, a fact we exploit extensively in section 5 when estimating the Hausdorff dimension of the free boundary of variational solutions.

For these variational solutions we derive a monotonicity formula which allows us to characterize the blow-up limits of variational solutions as backward self-similar functions (section 4).

This we apply to the following two problems:

(1) The quenching problem with respect to exponent $\gamma \in (-\frac{1}{3}, 0)$ (see, for example, [Ph] and [Le]):

$$\begin{gathered} {u_1}^{1+\gamma} \in L^1((0,\infty)\times \mathbf{R}^n), \\ u_1(0,x) = u_1^0(x),\ 0 \le u_1^0 \in C_0^{2,\alpha}(\mathbf{R}^n), \\ \partial_t u_1 - \Delta u_1 = -{u_1}^{\gamma}\, \chi_{\{u_1>0\}} \text{ in } (0,\infty)\times\mathbf{R}^n \end{gathered}$$

and u_1 is a solution in the sense of distributions obtained by the regularization of Phillips [Ph] which arises in chemical engineering (see [Ph]).

(2) The heat equation with Bernoulli-type boundary condition on the free boundary,

$$\begin{gathered} u_2 \ge 0,\ \partial_t u_2 - \Delta u_2 = 0 \text{ in } ((0,\infty)\times\mathbf{R}^n)\cap\{u_2>0\}, \\ |\nabla u_2| = 1 \text{ on } ((0,\infty)\times\mathbf{R}^n)\cap\partial\{u_2>0\}, \end{gathered} \tag{1.2}$$

which has been introduced as a model for flame propagation in [CaVa]. Here u_2 denotes a variational solution as introduced in Definition 3.1.

We show that the blow-up limits of u_1 and u_2 are backward self-similar variational solutions. Then by a standard dimension reduction procedure it is possible to estimate the Hausdorff dimension of special sets which we carry out only for the case of free boundaries: the Hausdorff dimension (with respect to the parabolic metric) of $\partial\{u_i > 0\}$ $(i = 1, 2)$ does not exceed $n+1$ and this estimate is optimal (Theorem 5.2).

2. Notation. Considering functions $u \in H^{1,2}_{\mathrm{loc}}(\mathbf{R}^n)$, $v \in H^{1,2}_{\mathrm{loc}}((0,T)\times\mathbf{R}^n)$, $\phi \in H^{1,2}_{\mathrm{loc}}(\mathbf{R}^n;\mathbf{R}^n)$, and $\psi \in H^{1,2}_{\mathrm{loc}}((0,T)\times\mathbf{R}^n;\mathbf{R}^n)$, we denote by $\nabla u := (\partial_1 u, \dots, \partial_n u)$ and $\nabla v := (\partial_2 v, \dots, \partial_{n+1} v)$ the space gradient, by $\partial_t v := \partial_1 v$ the time derivative, by $\operatorname{div}\phi := \sum_{i=1}^n \partial_i \phi_i$ and $\operatorname{div}\psi := \sum_{i=2}^{n+1} \partial_i \psi_i$ the space divergences, by $\nabla_{t,x} v := (\partial_1 v, \dots, \partial_{n+1} v)$ the time-space gradient, by $\operatorname{div}_{t,x}\psi := \sum_{i=1}^{n+1} \partial_i \psi_i$ the time-space divergence, and by

$$D\phi := \begin{pmatrix} \partial_1\phi_1 & \dots & \partial_n\phi_1 \\ & \dots & \\ \partial_1\phi_n & \dots & \partial_n\phi_n \end{pmatrix} \quad \text{and } D\psi := \begin{pmatrix} \partial_2\psi_1 & \dots & \partial_{n+1}\psi_1 \\ & \dots & \\ \partial_2\psi_{n+1} & \dots & \partial_{n+1}\psi_{n+1} \end{pmatrix}$$

the space Jacobian.

Moreover let us denote by χ_A the characteristic function of the set A, by $x\cdot y$ the Euclidean inner product in $\mathbf{R}^n\times\mathbf{R}^n$, by $|x|$ the Euclidean norm in $\mathbf{R}^n$, by $B_r(x_0) :=$

$\{x \in \mathbf{R}^n |\ |x-x_0| < r\}$ the ball of center x_0 and radius r, by $Q_r(t_0,x_0) := (t_0-r^2, t_0+r^2)\times B_r(x_0)$ the cylinder of radius r and height $2r^2$, by $T_r^-(t_0) := (t_0-4r^2, t_0-r^2)\times\mathbf{R}^n$ the horizontal layer from t_0-4r^2 to t_0-r^2 , by $T_r^+(t_0) := (t_0+r^2, t_0+4r^2)\times\mathbf{R}^n$ the horizontal layer from t_0+r^2 to t_0+4r^2, and by

$$G_{(t_0,x_0)}(t,x) := 4\pi(t_0-t)\,|4\pi(t_0-t)|^{-\frac{n}{2}-1}\ \exp\left(-\frac{|x-x_0|^2}{4(t_0-t)}\right)$$

the backward heat kernel, defined in $((-\infty,t_0)\cup(t_0,+\infty))\times\mathbf{R}^n$.

Furthermore, by ν we will always refer to the outer normal on a given surface. Finally $\mathcal{L}^n$ shall denote the n-dimensional Lebesgue measure, $\mathcal{H}^s$ the s-dimensional Hausdorff measure, and $\mathbf{W}_p^{2,1}$ and $\mathbf{H}^{1,\frac{1}{2}}$ the parabolic Sobolev- and Hölder-spaces as defined in [LSU].

3. Notion of solution and monotonicity formula. We begin by introducing our notion of a variational solution of the equation

$$\begin{gathered}\partial_t u-\Delta u = -\tfrac{\lambda_+}{2}\,p\,u^{p-1}\,\chi_{\{u>0\}}\ +\ \tfrac{\lambda_-}{2}\,p\,(-u)^{p-1}\,\chi_{\{u<0\}}\\ \text{in the case } p\in(0,2) \text{ and}\\ \partial_t u-\Delta u = 0 \text{ in } \{u>0\}\cup\{u<0\}\ ,\\ |\nabla\max(u,0)|^2-|\nabla\min(u,0)|^2 = \lambda_+-\lambda_- \text{ on } \partial\{u>0\}\cup\partial\{u<0\}\\ \text{in the case } p=0,\end{gathered}\tag{3.1}$$

where λ_+,λ_- are parameters $\in\mathbf{R}$.

For $p\ge 1$, nonnegative λ_+,λ_-, and suitable initial data there exists a unique strong solution of (3.1) and the proofs to follow can be readily simplified in this case.

DEFINITION 3.1. *We define $u\in H^{1,2}((t_1,t_2)\times B_R(0))$ for any $R\in(0,\infty)$ to be a* variational solution of (3.1) *if u is continuous in $(t_1,t_2)\times\mathbf{R}^n$ and continuously differentiable in the set $((t_1,t_2)\times\mathbf{R}^n)\cap(\{u>0\}\cup\{u<0\})$ once in time direction and twice in space directions and*

$$\begin{aligned}0=I :=& \int_{t_1+\delta}^{t_2-\delta}\int_{\mathbf{R}^n}\Big[\Big(|\nabla u|^2\ +\ \lambda_+\,\chi_{\{u>0\}}\,u^p\ +\ \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\Big)\ \operatorname{div}_{t,x}\psi\\ &-2\,\nabla_{t,x}u\,D\psi\,\nabla u\ -\ 2\,\partial_t u\,\nabla_{t,x}u\cdot\psi\Big]\\ &-\Big[\int_{\mathbf{R}^n}\Big(|\nabla u|^2\ +\ \lambda_+\,\chi_{\{u>0\}}\,u^p\ +\ \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\Big)\,\psi_1\Big]_{t_1+\delta}^{t_2-\delta}\\ =&\int_{t_1+\delta}^{t_2-\delta}\int_{\mathbf{R}^n}\Big[\Big(|\nabla u|^2\ +\ \lambda_+\,\chi_{\{u>0\}}\,u^p\ +\ \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\Big)\sum_{i=1}^{n+1}\partial_i\psi_i\\ &-2\sum_{j=2}^{n+1}\sum_{i=1}^{n+1}\partial_j u\,\partial_j\psi_i\,\partial_i u\ -\ 2\partial_t u\sum_{k=1}^{n+1}\partial_k u\,\psi_k\Big]\\ &-\int_{\mathbf{R}^n}\Big[\Big(|\nabla u|^2\ +\ \lambda_+\,\chi_{\{u>0\}}\,u^p\ +\ \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\Big)\,\psi_1\Big]\ (t_2-\delta)\end{aligned}$$

$$+ \int_{\mathbf{R}^n} \left[\left(|\nabla u|^2 \, + \, \lambda_+ \, \chi_{\{u>0\}} \, u^p \, + \, \lambda_- \, \chi_{\{u<0\}} \, (-u)^p \right) \psi_1 \right] \, (t_1 + \delta)$$

for almost every (a.e.) small and positive δ *and any* $\psi \in C^1(\mathbf{R}^{n+1})$ *such that* supp $\psi(t) \subset\subset \mathbf{R}^n$ *for any* $t \in (t_1, t_2)$. *Notice that for given* u, I *is formally the first variation with respect to variations of the domain* in time and space *of the functional*

$$\mathbf{G}(u, v) := \int_{t_1+\delta}^{t_2-\delta} F(v(t)) \, dt \, + \, \int_{t_1+\delta}^{t_2-\delta} \int_{\mathbf{R}^n} 2 \, \partial_t u \, v,$$

i.e., $I = -\frac{d}{d\epsilon} \mathbf{G}(u, u((t, x) + \epsilon\psi(t, x)))|_{\epsilon=0}$.

The continuity and differentiability assumptions on u are necessary in that they cannot be deduced from the other assumptions in Definition 3.1 by regularity theory, but they are rather mild in the sense that they can be verified without effort in many of the examples (see section 5). Existence of variational solutions and the relation to other notions of weak solutions will be discussed along with applications in section 5.

In this section we are going to derive a monotonicity formula for variational solutions with respect to any $p \in (-\infty, 2)$. In the case of continuous variational solutions with respect to $p \in [0, 2)$ this monotonicity will hold in arbitrary points of $(t_1, t_2) \times \mathbf{R}^n$, and in the case of strong solutions with respect to $p < 0$ it will hold as long as the solution continues to exist. So we exclude the quenching problem with a critical exponent,

$$\partial_t u - \Delta u = -\frac{1}{u} \, \chi_{\{u>0\}}.$$

THEOREM 3.1 (monotonicity formula). *Suppose that* $t_1 \le T \le t_2$, *that* $x_0 \in \mathbf{R}^n$, *that*

$$\sup_{t \in (t_1, T-\delta) \cup (T+\delta, t_2)} \int_{\mathbf{R}^n} \exp\left(-\frac{|x - x_0|^2}{4(T-t)} \right) \left(|\nabla u|^2 \, + \, |u|^p \right) (t, x)$$

$$+ \int_{(t_1, T-\delta) \cup (T+\delta, t_2)} \int_{\mathbf{R}^n} \exp\left(-\frac{|x - x_0|^2}{4(T-t)} \right) \left((\partial_t u)^2 \, + \, u^2 \right) (t, x) \, dx \, dt \, < \, \infty$$

for any positive δ, *and that either* $p \in [0, 2)$, u *is in* $((t_1, T) \cup (T, t_2)) \times \mathbf{R}^n$ *a variational solution in the sense of Definition* 3.1, *and* λ_+, λ_- *are nonnegative constants in the case* $p \in (0, 1)$, *or that* $p < 0$, $u \in C^0(((t_1, T) \cup (T, t_2)) \times \mathbf{R}^n)$, $u(t, x) \neq 0$ *in* $((t_1, T) \cup (T, t_2)) \times \mathbf{R}^n$, *and that* u *is a solution of* (3.1) *in the sense of distributions.*

Then for $\alpha = \frac{2}{2-p}$ *and* T_r^- , T_r^+, *and* $G_{(T,x_0)}$ *as defined in section* 2 *the functions*

$$\Psi^-(r) := r^{-2(\alpha-1)-2} \int_{T_r^-(T)} \left(|\nabla u|^2 \, + \, \lambda_+ \, \chi_{\{u>0\}} \, u^p \, + \, \lambda_- \, \chi_{\{u<0\}} \, (-u)^p \right) G_{(T,x_0)}$$

$$- \, \frac{\alpha}{2} \, r^{-2\alpha} \int_{T_r^-(T)} \frac{1}{T-t} \, u^2 \, G_{(T,x_0)}$$

and

$$\Psi^+(r) := r^{-2(\alpha-1)-2} \int_{T_r^+(T)} \left(|\nabla u|^2 \, + \, \lambda_+ \, \chi_{\{u>0\}} \, u^p \, + \, \lambda_- \, \chi_{\{u<0\}} \, (-u)^p \right) G_{(T,x_0)}$$

$$-\frac{\alpha}{2}\, r^{-2\alpha} \int_{T_r^+(T)} \frac{1}{T-t}\, u^2\, G_{(T,x_0)}$$

are well defined in the interval $(0, \frac{\sqrt{T-t_1}}{2})$ *and* $(0, \frac{\sqrt{t_2-T}}{2})$, *respectively, and satisfy for any* $0 < \rho < \sigma < \frac{\sqrt{T-t_1}}{2}$ *and* $0 < \rho < \sigma < \frac{\sqrt{t_2-T}}{2}$, *respectively, the monotonicity formulae*

$$\Psi^-(\sigma) \;-\; \Psi^-(\rho)$$

$$= \int_\rho^\sigma r^{-2\alpha-1} \int_{T_r^-(T)} \frac{1}{T-t}\, \left(\nabla u \cdot (x-x_0) \;-\; 2(T-t)\partial_t u \;-\alpha\, u\right)^2\, G_{(T,x_0)}\, dr$$

$$\geq 0$$

and

$$\Psi^+(\sigma) \;-\; \Psi^+(\rho)$$

$$= \int_\rho^\sigma r^{-2\alpha-1} \int_{T_r^+(T)} \frac{1}{T-t}\, \left(\nabla u \cdot (x-x_0) \;-\; 2(T-t)\partial_t u \;-\alpha\, u\right)^2\, G_{(T,x_0)}\, dr$$

$$\geq 0.$$

Remark 3.1. The integrand on the right-hand side of the monotonicity formula with respect to $x_0 = 0$ and $T = 0$,

$$-\left(\frac{1}{t}\right) \left(\nabla_{t,x} u \cdot (2t, x) \;-\; \alpha\, u(t,x)\right)^2\, G_{(T,x_0)}(t,x),$$

is sort of a measure of the distance of u to a function being homogeneous of degree α on paths $\theta \mapsto (\theta^2 t, \theta x)$. Since such a function would be of class C^β for any $\beta < \alpha$ on paths $\theta \mapsto (\theta^2 t, \theta x)$ but in general not of class C^α we may speak of a *monotonicity formula of order* α.

In the special case $\lambda_+ = \lambda_- = 0$ we obtain a monotonicity formula of order α for solutions of the heat equation which coincides when extrapolated at $\alpha = 0$ with the well-known monotonicity formula for the evolution of harmonic maps (see, for example, the proof of Theorem 8.1 of [St]).

Also, for $p > 2$ the integrands in our monotonicity formula coincide if written in similarity variables with those of the energy identity derived by Giga and Kohn in the classical paper [GiKo] for smooth solutions of (3.1); however, for $p \in [0, 1)$ globally smooth solutions no longer exist, and this is where the variational solutions of Definition 3.1 enter the game. The technique in our proof is motivated by that of Allard in section 5 of [Al], where he proves Fleming's monotonicity formula for stationary rectifiable n-varifolds.

Proof of Theorem 3.1. We give a proof for the monotonicity of Ψ^-. In order to obtain a proof with respect to Ψ^+ it is then sufficient to replace in what follows the interval $(-4r^2, -r^2)$ by $(r^2, 4r^2)$. Applying a translation we may assume that

$x_0 = 0$ and $T = 0$. Omitting the index $(0,0)$ of $G_{(0,0)}$ and $T_r^-(0,0)$ and choosing $t_1 := -4r^2$, $t_2 := -r^2$, and, after approximation, $\psi(t,x) := (2t,x)\, G(t,x)\, \eta_\kappa(x)$ in Definition 3.1 where $\eta_\kappa \in H_0^{1,\infty}(\mathbf{R}^n)$ we get for a.e. $r \in (0, \frac{\sqrt{T-t_1}}{2})$ the identity

$$
\begin{aligned}
0 = \int_{T_r^-} \Bigg[& \Big(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p \Big) \Big(2G + 2t\,\partial_t G + \,\operatorname{div}(xG) \Big) \eta_\kappa \\
& - 2\,\eta_\kappa \sum_{j=2}^{n+1} \sum_{i=2}^{n+1} \partial_j u \;(\delta_{ji}\, G \,+\, \partial_j G\, x_i)\; \partial_i u \;-\; 2\eta_\kappa \sum_{j=2}^{n+1} \partial_j u\, \partial_j G\, 2t\, \partial_t u \\
& - 2\eta_\kappa \sum_{j=2}^{n+1} \partial_j u\, G\, x_j\, \partial_t u \;-\; 2\eta_\kappa\, (\partial_t u)^2\, 2t\, G \Bigg] \\
& - \int_{\mathbf{R}^n} \Big[2t\,\eta_\kappa \Big(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p \Big)\, G \Big] (-r^2) \\
& + \int_{\mathbf{R}^n} \Big[2t\,\eta_\kappa \Big(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p \Big)\, G \Big] (-4r^2) \\
& + \int_{T_r^-} \Big(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p \Big)\, \nabla \eta_\kappa \cdot x\, G \\
& - 2 \sum_{j=2}^{n+1} \sum_{i=2}^{n+1} \partial_j u\, \partial_j \eta_\kappa\, x_i\, \partial_i u\, G \;-\; 2 \sum_{j=2}^{n+1} \partial_j u\, \partial_j \eta_\kappa\, 2t\, \partial_t u\, G.
\end{aligned}
$$

Next, we derive the identity

$$
\begin{aligned}
(3.2) \int_{T_r^-} |\nabla u|^2\, G\, \eta_\kappa \;=\; - \int_{T_r^-} \Big[& u\, \eta_\kappa\, \nabla u \cdot \nabla G \,+\, \eta_\kappa\, G \Big(u\, \partial_t u \,+\, \frac{p}{2} (\lambda_+ \,\chi_{\{u>0\}}\, u^p \\
& + \lambda_- \,\chi_{\{u<0\}}\, (-u)^p) \Big) \Big] \;-\; \int_{T_r^-} u\, G\, \nabla u \cdot \nabla \eta_\kappa.
\end{aligned}
$$

First notice that $\max(u,\theta)$ is for small positive θ a subsolution of

$$
-\partial_t \max(u,\theta) \,+\, \Delta \max(u,\theta) - \frac{\lambda_+}{2}\, p\, \chi_{\{u>\theta\}}\, u^{p-1} \geq 0
$$

and that the nonnegative distribution

$$
-\partial_t \max(u,\theta) \,+\, \Delta \max(u,\theta) - \frac{\lambda_+}{2}\, p\, \chi_{\{u>\theta\}}\, u^{p-1}
$$

is a σ-finite σ-additive measure with support in $\partial\{u > \theta\}$. Since $\frac{\lambda_+}{2}\, p\, \chi_{\{u>\theta\}}\, u^{p-1} \to \frac{\lambda_+}{2} p \chi_{\{u>0\}} u^{p-1}$ in $L^1_{\mathrm{loc}}((t_1,t_2)\times\mathbf{R}^n)$ and $\partial_t \max(u,\theta) \to \partial_t \max(u,0)$ in $L^2_{\mathrm{loc}}((t_1,t_2)\times$

$\mathbf{R}^n$) as $\theta \searrow 0$, we obtain that $\Delta \max(u,\theta) \rightharpoonup \Delta \max(u,0)$ weakly-* in the space $(C^0(\overline{(t_1+\delta,t_2-\delta)\times B_R(0)}))^\star$ for each $\delta > 0$ as $\theta \searrow 0$ and that

$$\text{(3.3)} \quad \operatorname{supp}(-\partial_t \max(u,0) \,+\, \Delta \max(u,0) - \frac{\lambda_+}{2}\, p\, \chi_{\{u>0\}}\, u^{p-1}) \subset \partial\{u>0\}\,.$$

Considering mollified functions approximating $\max(u,0)$, we now see that

$$\int \nabla \max(u,0)\cdot\nabla\zeta \;=\; -\int \zeta \Delta \max(u,0)$$

for any $\zeta \in C^0(\overline{(t_1+\delta,t_2-\delta)\times B_R(0)}) \cap L^2((t_1+\delta,t_2-\delta); H_0^{1,2}(B_R(0)))$. An analogous formula holds for $-\min(u,0)$. Using this and (3.3) one can now easily derive the formula (3.2).

Now, multiplying the identity at the beginning of the proof by $-r^{-3-2(\alpha-1)}$, choosing $\eta_\kappa(x) := \min(1,\max(0,2-\kappa|x|))$ for small positive κ, and using the fact that $\nabla G = \frac{x\,G}{2t}$ and $\partial_t G \,+\, \Delta G = 0$ in $\{t<0\}\cup\{t>0\}$ we obtain

$$0 = r^{-3-2(\alpha-1)} \left[\int_{\mathbf{R}^n} 2t\,\eta_\kappa \left(|\nabla u|^2 \,+\, \lambda_+\,\chi_{\{u>0\}}\,u^p \,+\, \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\right) G\right]_{-4r^2}^{-r^2}$$

$$+\,(-2(\alpha-1)-2)\,r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa \left(|\nabla u|^2 \,+\, \lambda_+\,\chi_{\{u>0\}}\,u^p\right.$$

$$\left. +\,\lambda_-\,\chi_{\{u<0\}}\,(-u)^p\right) G \,+\, (2(\alpha-1)+2-2+2)\,r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa\,|\nabla u|^2\,G$$

$$+\,(2(\alpha-1)+2-2)\,r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa \left(\lambda_+\,\chi_{\{u>0\}}\,u^p \,+\, \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\right) G$$

$$+\,r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa \left[2\nabla u\cdot\nabla G\,\nabla u\cdot x + 4t\,\nabla G\cdot\nabla u\,\partial_t u + 2\,\nabla u\cdot x\,G\,\partial_t u + 4t\,(\partial_t u)^2\,G\right]$$

$$-\,r^{-3-2(\alpha-1)} \int_{T_r^-} \left[\left(|\nabla u|^2 \,+\, \lambda_+\,\chi_{\{u>0\}}\,u^p \,+\, \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\right) G\,\nabla\eta_\kappa\cdot x\right.$$

$$\left. -\,2\nabla u\cdot x\,\nabla u\cdot\nabla\eta_\kappa\,G \;-\; 4t\,\nabla u\cdot\nabla\eta_\kappa\,G\,\partial_t u\right]$$

$$= r^{-3-2(\alpha-1)} \left[\int_{\mathbf{R}^n} 2t\,\eta_\kappa \left(|\nabla u|^2 \,+\, \lambda_+\,\chi_{\{u>0\}}\,u^p \,+\, \lambda_-\,\chi_{\{u<0\}}\,(-u)^p\right) G\right]_{-4r^2}^{-r^2}$$

$$+\,(-2(\alpha-1)-2)\,r^{-3-2(\alpha-1)} \int_{T_r^-} \left(|\nabla u|^2 \,+\lambda_+\,\chi_{\{u>0\}}u^p \,+\lambda_-\,\chi_{\{u<0\}}(-u)^p\right) G\,\eta_\kappa$$

$$+ (2(\alpha-1) \,-\, \alpha\, p)\, r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa \left(\lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p\right) G$$

$$- r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa\, G \left(\frac{\alpha}{t}\, u\, \nabla u\cdot x + 2\alpha\, u\, \partial_t u - \frac{1}{t}\,(\nabla u\cdot x)^2 - 4\nabla u\cdot x\, \partial_t u - 4t\,(\partial_t u)^2\right)$$

$$- r^{-3-2(\alpha-1)} \int_{T_r^-} \Big[\left(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p\right) G\, \nabla\eta_\kappa\cdot x$$

$$- 2\nabla u\cdot x\, \nabla u\cdot\nabla\eta_\kappa\, G \,-\, 4t\, \nabla u\cdot\nabla\eta_\kappa\, G\, \partial_t u\Big]$$

$$= r^{-3-2(\alpha-1)} \left[\int_{\mathbf{R}^n} 2t\, \eta_\kappa \,\left(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p \,+\, \lambda_- \,\chi_{\{u<0\}}\, (-u)^p\right)\, G\right]_{-4r^2}^{-r^2}$$

$$+ (-2(\alpha-1)-2)\, r^{-3-2(\alpha-1)} \int_{T_r^-} \eta_\kappa \left(|\nabla u|^2 \,+\, \lambda_+ \,\chi_{\{u>0\}}\, u^p\right.$$

$$\left. + \lambda_- \,\chi_{\{u<0\}}\, (-u)^p\right) G \,- r^{-2\alpha-1} \int_{T_r^-} \frac{\eta_\kappa G}{(-t)}\, (\nabla u\cdot x \,+\, 2t\, \partial_t u \,-\, \alpha\, u)^2$$

$$+ r^{-2\alpha-1} \int_{T_r^-} \eta_\kappa\, G \left(\frac{\alpha}{t}\, u\, \nabla u\cdot x \,+\, 2\alpha\, u\, \partial_t u \,-\, \frac{\alpha^2}{t}\, u^2\right) \,+\, o(1) \quad \text{as } \kappa\to 0.$$

Realizing finally that

$$r^{-2\alpha-1} \int_{T_r^-} \eta_\kappa\, G \left(\frac{\alpha}{t}\, u\, \nabla u\cdot x \,+\, 2\alpha\, u\, \partial_t u \,-\, \frac{\alpha^2}{t}\, u^2\right)$$

$$= \alpha \int_{T_1^-} \eta_\kappa(rx)\, \frac{G(t,x)}{t}\, r^{-\alpha}\, u(r^2 t, rx) \left(r^{-\alpha}\, (\nabla u)(r^2 t, rx)\cdot x\right.$$

$$\left. + r^{-\alpha+1}\, 2t\, (\partial_t u)(r^2 t, rx) \,-\, \alpha\, r^{-\alpha-1}\, u(r^2 t, rx)\right)$$

$$= o(1) \,+\, \partial_r \left(\frac{\alpha}{2} \int_{T_1^-} \left(\frac{u(r^2 t, rx)}{r^\alpha}\right)^2 \,\frac{G(t,x)}{t}\, \eta_\kappa(rx)\right)$$

$$= o(1) \,+\, \partial_r \left(-\frac{\alpha}{2} \int_{T_r^-} \left(\frac{u}{r^\alpha}\right)^2 \,\frac{G}{(-t)}\, \eta_\kappa\right) \quad \text{as } \kappa\to 0,$$

integrating the resulting identity from ρ to σ, using integration by parts, and letting $\kappa\to 0$ the theorem is proved. □

4. Characterization of blow-up limits. The result to be presented in this section is that assuming some regularity of the solution in "parabolic direction" $(2\theta t, x)$ on the path $\theta \mapsto (T+\theta^2 t, x_0+\theta x)$, it is possible to prove that the solution's derivatives at (T, x_0) in this direction vanish to a certain order; this implies, of course, that any blow-up limit at (T, x_0) is a self-similar solution. The required regularity takes the simple form of a growth estimate.

THEOREM 4.1. *Suppose that* $t_1 \le T \le t_2$, *that* $x_0 \in \mathbf{R}^n$, *that*

$$\sup_{t\in(t_1,T-\delta)\cup(T+\delta,t_2)} \int_{\mathbf{R}^n} \exp\left(-\frac{|x-x_0|^2}{4(T-t)}\right)\left(|\nabla u|^2 + |u|^p\right)(t)$$

$$+ \int_{(t_1,T-\delta)\cup(T+\delta,t_2)} \int_{\mathbf{R}^n} \exp\left(-\frac{|x-x_0|^2}{4(T-t)}\right)\left((\partial_t u)^2 + u^2\right) < \infty$$

for any positive δ, *that* $0 < \rho_k \to 0$ *as* $k \to \infty$, *and that either*

(i) $p \in [0,2)$, u *is in* $((t_1,T)\cup(T,t_2)) \times \mathbf{R}^n$ *a variational solution in the sense of Definition* 3.1 *and* λ_+, λ_- *are nonnegative constants in the case* $p \in (0,1)$ *or*

(ii) $p < 0$, $u \in C^0(((t_1,T)\cup(T,t_2)) \times \mathbf{R}^n)$, $u \neq 0$ *in* $((t_1,T)\cup(T,t_2)) \times \mathbf{R}^n$, *and* u *is a solution of* (3.1) *in the sense of distributions.*

Suppose furthermore that in either case the growth estimates

$$\sup_{r\in(0,\frac{\sqrt{T-t_1}}{4})} \max\left(r^{-2\alpha}\int_{T_r^-(T)} \frac{1}{T-t}\, u^2\, G_{(T,x_0)}\right.$$

$$- r^{-2(\alpha-1)-2}\int_{T_r^-(T)} \lambda_+\, \chi_{\{u>0\}}\, u^p\, G_{(T,x_0)},$$

$$\left. - r^{-2(\alpha-1)-2}\int_{T_r^-(T)} \lambda_-\chi_{\{u<0\}}(-u)^p G_{(T,x_0)}\right) < +\infty$$

and

$$\sup_{r\in(0,\frac{\sqrt{t_2-T}}{4})} \max\left(r^{-2\alpha}\int_{T_r^+(T)} \frac{1}{T-t}\, u^2\, G_{(T,x_0)},\ - r^{-2(\alpha-1)-2}\int_{T_r^+(T)} |\nabla u|^2\, G_{(T,x_0)}\right.,$$

$$- r^{-2(\alpha-1)-2}\int_{T_r^+(T)} \lambda_+\, \chi_{\{u>0\}}\, u^p\, G_{(T,x_0)}\,,$$

$$\left. - r^{-2(\alpha-1)-2}\int_{T_r^+(T)} \lambda_-\, \chi_{\{u<0\}}\, (-u)^p\, G_{(T,x_0)}\right) < +\infty$$

are satisfied. Then $\Psi^-(r) \searrow M^-(u,(T,x_0))$ *as* $r \searrow 0$ *provided that* $T > t_1$ *and* $\Psi^+(r) \searrow M^+(u,(T,x_0))$ *as* $r \searrow 0$ *provided that* $T < t_2$, *and for any* $D \subset\subset (((-\infty)\sqrt{T-t_1}, 0) \cup (0, ((+\infty)\sqrt{t_2-T}))) \times \mathbf{R}^n$ *and* $k \ge k(D)$ *the sequence*

$$u_k(t,x) := \frac{u(T+{\rho_k}^2 t, x_0+\rho_k x)}{{\rho_k}^\alpha}$$

is bounded in $H^{1,2}(D) \cap L^p(D)$ *and any weak* $H^{1,2}$*-limit* u_0 *with respect to a subsequence is a function homogeneous of degree* α *on paths* $\theta \mapsto (\theta^2 t, \theta x)$ *for* $\theta > 0$ *and* $(t,x) \in (((-\infty)\sqrt{T-t_1}, 0) \cup (0, ((+\infty)\sqrt{t_2 - T}))) \times \mathbf{R}^n$, *i.e.,*

$$u_0(\lambda^2 t, \lambda x) = \lambda^\alpha \, u_0(t,x) \quad \textit{for any } \lambda > 0$$

$$\textit{and for any } (t,x) \in (((-\infty)\sqrt{T-t_1}, 0) \cup (0, ((+\infty)\sqrt{t_2 - T}))) \times \mathbf{R}^n.$$

Proof. To avoid clumsy notation we give the proof only for the case $t_2 = T$. First we calculate for $0 < R < S < \infty$ (to be used later)

$$\Psi^-(\rho_k R) = R^{-2(\alpha-1)-2} \int_{T_R^-(0)} \left(|\nabla u_k|^2 + \lambda_+ \, \chi_{\{u_k > 0\}} \, {u_k}^p + \lambda_- \, \chi_{\{u_k < 0\}} \, (-u_k)^p \right) G_{(0,0)}$$

$$- \frac{\alpha}{2} \, R^{-2\alpha} \int_{T_R^-(0)} \frac{1}{(-t)} \, {u_k}^2 \, G_{(0,0)}.$$

Hence, using the assumed growth estimate, the monotonicity formula Theorem 3.1 yields that for $k \geq k(D)$ the sequences u_k and ∇u_k are bounded in $L^2(D)$ and u_k is bounded in $L^p(D)$.

Since Ψ^- is nondecreasing and bounded in $(0, r_0)$ for small positive r_0, we know that Ψ^- has a real right limit at 0. Consequently,

$$0 \leftarrow \Psi^-(\rho_k S) - \Psi^-(\rho_k R)$$

$$= \int_R^S r^{-2\alpha-1} \int_{T_r^-(0)} \frac{1}{(-t)} \, (\nabla u_k \cdot x \, + \, 2t \partial_t u_k \, - \alpha \, u_k)^2 \, G_{(0,0)} \, dr$$

as $k \to \infty$. Thus for $k \geq k(D)$ the sequence u_k is bounded in $H^{1,2}(D)$, and passing to a subsequence $k \to \infty$ such that $u_k \rightharpoonup u_0$ weakly in $H^{1,2}_{\mathrm{loc}}((-\infty, 0) \times \mathbf{R}^n)$ and using the lower semicontinuity of the L^2-norm with respect to weak convergence we obtain $\nabla u_0(t,x) \cdot x + 2t \partial_t u_0(t,x) - \alpha u_0(t,x) = 0$ a.e. in $(-\infty, 0) \times \mathbf{R}^n$. Now it is easily seen that u_0 is homogeneous of degree α on paths $\theta \mapsto (\theta^2 t, \theta x)$ for $\theta > 0$ and $(t,x) \in (-\infty, 0) \times \mathbf{R}^n$. □

5. Applications. In this section we will characterize the blow-up limits with respect to two one-phase problems, the quenching problem with exponent $\gamma \in (-\frac{1}{3}, 0)$ and the heat equation with a Bernoulli type boundary condition on the free boundary.

Before starting with these two equations let us briefly consider solutions of the heat equation with strong absorption, i.e., $\gamma \in [0,1), u \in L^{1+\gamma}((0,\infty) \times \mathbf{R}^n) \cap L^\infty((0,\infty) \times \mathbf{R}^n)$ satisfying $\partial_t u - \Delta u = -u^\gamma \chi_{\{u>0\}}$ in $(0,\infty) \times \mathbf{R}^n$, $u(0,x) = u^0(x)$ in the sense of distributions. Here we assume $0 \leq u^0 \in C_0^{2,\alpha}(\mathbf{R}^n)$ for some $\alpha \in (0,1)$ and $(u^0)^{-\gamma} \Delta u^0 \in L^\infty(\mathbf{R}^n)$.

This problem, in particular the solution's behavior at extinction points, has been extensively studied (see [FrHe] to cite one study). For issues not concentrating on the extinction angle see, for example, [Ca] and [ChWe].

Since we assume u to be bounded, the comparison principle holds and it follows immediately from L^p-estimates that the now unique $u \in \mathbf{W}^{2,1}_p((\frac{1}{R}, R) \times B_R(0))$ for

any $0 < R < \infty$ and $1 \le p < \infty$. Comparing u to $\max(0, (1-\gamma)(\frac{(\sup_{\mathbf{R}^n} u^0)^{1-\gamma}}{1-\gamma} - t))^{\frac{1}{1-\gamma}}$, we see that $u(t,x) = 0$ for $t \ge \frac{(\sup_{\mathbf{R}^n} u^0)^{1-\gamma}}{1-\gamma}$; furthermore, standard energy estimates imply that $\nabla u \in L^\infty((0,\infty); L^2(\mathbf{R}^n))$ and $\partial_t u \in L^2((0,\infty)\times\mathbf{R}^n)$. Thus,

$$\int_0^T \int_{\mathbf{R}^n} \left(|\nabla u|^2 \, + \, \frac{2}{1+\gamma} u^{1+\gamma} \, + \, 2\partial_t u \, u \right)$$

$$- \int_0^T \int_{\mathbf{R}^n} \left(|\nabla v|^2 \, + \, \frac{2}{1+\gamma} v^{1+\gamma} \, + \, 2\partial_t u \, v \right)$$

$$\le \int_0^T \int_{\mathbf{R}^n} \left(-|\nabla(u-v)|^2 \, + \, 2\nabla u \cdot \nabla(u-v) \, + \, 2u^\gamma \, (u-v) \, + \, 2\partial_t u \, (u-v) \right) \, \le \, 0$$

for any $v \in L^2((0,T); H^{1,2}_{\mathrm{loc}}(\mathbf{R}^n))$ such that $\mathrm{supp}(v-u)(t) \subset\subset B_R(0)$ for any $t \in (0,T)$. Consequently u satisfies the variational inequality

$\mathbf{G}(u,u) \le \mathbf{G}(u,v)$ for any such v and the $\mathbf{G}$ in Definition 3.1, and it follows that u is in $(0,\infty)\times\mathbf{R}^n$ a variational solution in the sense of Definition 3.1.

In order to be allowed to apply Theorem 4.1 in the past of any point $(T, x_0) \in \partial\{u > 0\} \cap ((0,\infty)\times\mathbf{R}^n)$, it is therefore sufficient to verify the growth estimate

$$\sup_{r \in (0, \frac{\sqrt{T}}{4})} r^{-\frac{4}{1-\gamma}} \int_{T_r^-(T)} \frac{1}{T-t} \, u^2 \, G_{(T,x_0)} \, < \, +\infty,$$

but this follows directly from the regularity estimate Theorem 3.1 derived in [ChWe],

$$\| u^{\frac{1-\gamma}{2}} \|_{\mathbf{H}^{1,\frac{1}{2}}((\delta,\frac{1}{\delta})\times B_{\frac{1}{\delta}}(0))} \, \le \, C(\delta), \tag{5.1}$$

and from the fact that $u \in L^\infty((0,\infty)\times\mathbf{R}^n)$. Therefore any blow-up limit of u must be a backward self-similar solution. Let us conclude this short observation with the remark that there exists a variety of such self-similar solutions; for example,

$$v_{k,\sigma}(t,x) := -\sigma \, t \, + \, (1-\sigma) \, \frac{1}{2k} \, \sum_{i=1}^k {x_i}^2 \quad \text{and}$$

$$w_\sigma(t,x) := \frac{1}{2} \, \max(0, x_1)^2$$

are for $t < 0$, $k \in \{1, \dots, n\}$, and $\sigma \in [0,1]$ self-similar solutions with respect to $\gamma = 0$. Notice that, of the stationary solutions ($\sigma = 0$), only the half-plane solution $\frac{1}{2} \, \max(0, x_1)^2$ has a regular free boundary.

We proceed to the quenching problem: Here we consider certain distributional solutions u of

$$\begin{gathered} u^{1+\gamma} \in L^1((0,\infty) \times \mathbf{R}^n), \; \partial_t u - \Delta u = -u^\gamma \, \chi_{\{u>0\}} \text{ in } (0,\infty)\times\mathbf{R}^n, \\ u(0,x) = u^0(x), \\ \gamma \in (-1,0) \text{ and } 0 \le u^0 \in C_0^{2,\alpha}(\mathbf{R}^n) \text{ for some } \alpha \in (0,1). \end{gathered} \tag{5.2}$$

Existence of a distributional solution u of (5.2) as well as optimal regularity in space has been shown by Phillips (see [Ph]) by way of a regularization of the equation in (5.2):

$$\text{for } f_\epsilon(z) := \begin{cases} 0, & z \le 0, \\ \frac{z}{\epsilon + z^{1-\gamma}}, & z > 0; \end{cases}$$

the equation $\partial_t v - \Delta v = -f_\epsilon(v)$ in $(0,\infty)\times\mathbf{R}^n$, $v(0,x) = u^0_\epsilon(x)$, $0 \le u^0_\epsilon \in C^\infty_0(\mathbf{R}^n)$, $u^0_\epsilon \to u^0$ in $C^{2,\alpha}(\mathbf{R}^n)$ as $\epsilon \to 0$ admits for positive ϵ a unique solution $0 \le u_\epsilon \in C^0_0([0,\infty)\times\mathbf{R}^n) \cap C^2((0,\infty)\times\mathbf{R}^n)$ such that

$$\nabla\left(u_\epsilon^{\frac{1-\gamma}{2}}\right) \text{ is bounded in } L^\infty((0,\infty)\times\mathbf{R}^n),$$

u_ϵ is bounded in $C^{0,\frac{1}{3n}}([0,\infty)\times\mathbf{R}^n)$, and as a subsequence $\epsilon \to 0$ the solutions u_ϵ converge in $C^0([0,\infty)\times\mathbf{R}^n)$ to a nonnegative solution u of (5.2) in the sense of distributions.

In order to verify the growth assumptions in Theorem 4.1 it is necessary to complete the optimal regularity of u in space by the optimal regularity estimate in time:

$$\|\partial_t(u^{1-\gamma})\|_{L^\infty((\delta,\frac{1}{\delta})\times B_{\frac{1}{\delta}}(0))} \;\le\; C(\delta).$$

For proof observe first that $u^{1-\gamma}$ satisfies the equation

$$\partial_t(u^{1-\gamma}) - \Delta(u^{1-\gamma}) = -\chi_{\{u^{1-\gamma}>0\}}\left((1-\gamma) - \frac{4\gamma}{1-\gamma}|\nabla(u^{\frac{1-\gamma}{2}})|^2\right)$$

in $(0,\infty)\times\mathbf{R}^n$ which may be written as

$$\partial_t w - \Delta w = -\chi_{\{w>0\}}\, g \text{ with } 0 \le w := u^{1-\gamma} \text{ and } g \in L^\infty((0,\infty)\times\mathbf{R}^n)\,.$$

LEMMA 5.1. *There exists a constant $C < \infty$ depending on $\|g\|_{L^\infty((0,\infty)\times\mathbf{R}^n)}$ and n such that any nonnegative solution $w \in L^1((0,\infty)\times\mathbf{R}^n)$ of $\partial_t w - \Delta w = -g\,\chi_{\{w>0\}}$ in $(0,\infty)\times\mathbf{R}^n$ satisfies for any $Q_{2r}(t_0,x_0) \subset (0,\infty)\times\mathbf{R}^n$ the estimate*

$$\sup_{P_r(t_0,x_0)} w \;\le\; C\,(w(t_0,x_0) \,+\, r^2)$$

(here $P_r(t_0,x_0) := \{(t,x) : 0 > t - t_0 > -r^2$ and $|x-x_0|^2 < t_0 - t\}$).

Proof. We observe that the function $w_r(t,x) := r^{-2}\, w(t_0 + r^2 t, x_0 + rx)$ satisfies in $Q_2(0) = (-4,4)\times B_2(0)$ the equation $\partial_t w_r - \Delta w_r = -g_r\,\chi_{\{w_r>0\}}$ where $g_r(t,x) := g(t_0 + r^2 t, x_0 + rx)$. Therefore we may apply Harnack's inequality to obtain a constant $C_1 < \infty$ depending only on n and $\|g\|_{L^\infty}$ such that

$$\sup_{(-1,-\frac{1}{2})\times B_1(0)} w_r \le C_1\,(w_r(0) \,+\, 1).$$

Scaling back yields

$$\sup_{P_r(t_0,x_0)} w \;\le\; C\,(w(t_0,x_0) \,+\, r^2). \qquad \square$$

LEMMA 5.2. *For the solution u from above and any positive δ there exists $C < \infty$ such that the function $w = u^{1-\gamma}$ satisfies*

$$\|\partial_t w\|_{L^\infty((\delta,\infty)\times\mathbf{R}^n)} \le C.$$

Proof. Suppose that there exists $(t_k, x_k) \in \{w > 0\} \cap ((\delta,\infty)\times\mathbf{R}^n)$ such that $|\partial_t w|(t_k,x_k) \to +\infty$ as $k \to \infty$: since $w \in C_0^0([0,\infty)\times\mathbf{R}^n)$ we conclude that $(t_k,x_k) \to (t_0,x_0)$ as a subsequence $k \to \infty$. Then we consider the sequence $w_k(t,x) := {\rho_k}^{-2} w(t_k + {\rho_k}^2 t, x_k + \rho_k x)$ where $\rho_k := w(t_k,x_k)^{\frac{1}{2}}$ can be assumed to go to 0 as $k \to \infty$. From Phillips' space gradient estimate we now know that $w_k(0,x) \le 2 + C_1|x|^2$ and Lemma 5.1 implies that $\sup_{P_1(0,x)} w_k \le C_2(2+|x|^2)$. Thus w_k is bounded in $L^\infty(Q_1^-(0))$ (where $Q_r^-(0) = (-r^2,0)\times B_r(0)$) and we infer from parabolic regularity theory that w_k is bounded in $C^{0,\beta}([-\frac{1}{2},0]\times\overline{B_{\frac{1}{2}}(0)})$ for $\beta \in (0,1)$. Consequently there is a subsequence $k \to \infty$ such that $w_k \to w_0$ in $C^0([-\frac{1}{2},0]\times\overline{B_1(0)})$, and the fact that $w_k(0) = 1$ implies that $w_k \ge c > 0$ in $\overline{Q_\kappa^-(0)}$ for some $\kappa > 0$ and large k of the subsequence.

But then $u_k(t,x) := {\rho_k}^{-\frac{2}{1-\gamma}} u(t_k + {\rho_k}^2 t, x_k + \rho_k x) \ge c^{\frac{1}{1-\gamma}} > 0$ in $\overline{Q_\kappa^-(0)}$ and u_k is bounded in $L^\infty(Q_1^-(0))$, and we obtain from regularity theory for the equation of u_k that u_k is for large k uniformly positive and bounded in $C^m(\overline{Q_{\frac{\kappa}{2}}^-(0)})$ for $m < \infty$, a contradiction to the assumption that $+\infty \leftarrow |\partial_t w|(t_k,x_k) = (1-\gamma)u^{-\gamma}(t_k,x_k)|\partial_t u(t_k,x_k)| = (1-\gamma){u_k}^{-\gamma}(0)|\partial_t u_k(0)|$. □

We may also state the corresponding nondegeneracy estimate.

LEMMA 5.3. *For the solution u from above and any $\delta > 0$ there exists $c > 0$ such that for any $(T,x_0) \in \overline{\{u>0\}}$ and any $Q_r^-(T,x_0) \subset ((\delta,\infty)\times\mathbf{R}^n)$ the estimate*

$$\sup_{Q_r^-(T,x_0)} u \ge c\, r^{\frac{2}{1-\gamma}}$$

holds.

Proof. Notice that it is sufficient to prove the existence of c depending only on n and γ such that the estimate is satisfied in points of $\{u > 0\}$. Since we have $\partial_t(u^{1-\gamma}) - \Delta(u^{1-\gamma}) \le -\chi_{\{u^{1-\gamma}>0\}}$, $w := u^{1-\gamma}$ is a subsolution of the parabolic obstacle problem.

As before we use scaled functions $w_r(t,x) := r^{-2} w(t_0 + r^2 t, x_0 + rx)$. Furthermore, we introduce a comparison function

$$v(t,x) := \frac{1}{4n}\,(|x|^2 - 2n\,t);$$

obviously v solves $\partial_t v - \Delta v = -1$ in $Q_1^-(0)$, and of course v is positive in $Q_1^-(0)$, it is positive on $(\{-1\}\times\overline{B_1(0)}) \cup ([-1,0]\times\partial B_1(0))$, and it vanishes in the point 0.

Now, if $w_r \le v$ on the *whole* parabolic boundary of $Q_1^-(0)$, then the comparison principle would imply $w_r \le v$ in $Q_1^-(0)$, a contradiction to the assumption $w_r(0) > 0$. Therefore there has to be a point on the parabolic boundary of $Q_1^-(0)$ in which $w_r > v$, and scaling back proves the statement of our proposition. □

Finally we have to verify that u is a variational solution in the sense of Definition 3.1. To this end, we confine ourselves to the case $\gamma \in (-\frac{1}{3},0)$ and start with another nondegeneracy estimate whose proof follows Lemma 7.3 of [ChWe].

LEMMA 5.4. *For the solution u from above and any $p \in (0, \frac{1-\gamma}{2})$ there exist $r_0 > 0$ and $C < \infty$ such that for any $(t_0, x_0) \in \partial\{u > 0\} \cap ((\delta, \infty) \times \mathbf{R}^n)$ and any $r \leq r_0$*

$$\int_{Q_r(t_0,x_0)\cap\{u>0\}} u^{-p} \;\leq\; C\, r^{n+2-\frac{2p}{1-\gamma}} \quad .$$

Proof. Let us define $w := u^{1-\gamma-p}$, which satisfies then in $\{u > 0\}$ the equation $\partial_t w - \Delta w = -(1-\gamma-p)u^{-p}\big(1-(\gamma+p)\frac{|\nabla u|^2}{u^{1+\gamma}}\big)$. Hence $\partial_t w - \Delta w \leq -(1-\gamma-p)u^{-p}\big(1-\frac{2(1+\eta)(\gamma+p)}{1+\gamma}\big) \leq -\kappa\, u^{-p} < 0$ in $Q_r(t_0,x_0)\cap\{u>0\}$ by Lemma 2 of [Ph], provided that η and r_0 have been chosen small enough. Furthermore we know from Lemma 2 of [Ph] that there exist constants C_1, C_2, C_3 such that $|\nabla w| \;\leq\; C_1\, u^{\frac{1+\gamma}{2}-\gamma-p} \;\leq\; C_2\, r^{1-\frac{2p}{1-\gamma}}$ and that $w \;\leq\; C_3\, r^{\frac{2(1-\gamma-p)}{1-\gamma}}$ in $Q_r(t_0,x_0)$.

Introducing now a C^∞-function ρ satisfying $\rho \geq 0$, $\rho' \geq 0$, $\rho'' \geq 0$ as well as $\rho(\tau) = \tau - 1$, $\tau \geq 2$ and $\rho(\tau) = 0$, $\tau \leq \frac{1}{2}$ and defining $\rho_\delta(\tau) := \delta\rho(\frac{\tau}{\delta})$ we obtain for $0 < \delta << \epsilon$

$$0 \leq \frac{1}{\epsilon}\int_{Q_r(t_0,x_0)} \nabla w \cdot \nabla \min(\epsilon, \rho_\delta(u)) \;=\; -\frac{1}{\epsilon}\int_{Q_r(t_0,x_0)\cap\{0<\rho_\delta(u)\leq\epsilon\}} \rho_\delta(u)\,\Delta w$$

$$-\int_{Q_r(t_0,x_0)\cap\{\rho_\delta(u)>\epsilon\}} \Delta w \;+\; \int_{t_0-r^2}^{t_0+r^2}\int_{\partial B_r(x_0)} \frac{\min(\epsilon,\rho_\delta(u))}{\epsilon}\,\nabla w\cdot\nu\, d\mathcal{H}^{n-1}.$$

Consequently

$$\kappa\int_{Q_r(t_0,x_0)\cap\{\rho_\delta(u)>\epsilon\}} u^{-p} \;\leq\; C_4\, r^{n+2-\frac{2p}{1-\gamma}} \;-\; \frac{\kappa}{\epsilon}\int_{Q_r(t_0,x_0)\cap\{0<\rho_\delta(u)\leq\epsilon\}} \rho_\delta(u)\, u^{-p}$$

$$-\frac{1}{\epsilon}\int_{Q_r(t_0,x_0)\cap\{0<\rho_\delta(u)\leq\epsilon\}} \rho_\delta(u)\,\partial_t w \;-\; \int_{Q_r(t_0,x_0)\cap\{\rho_\delta(u)>\epsilon\}} \partial_t w,$$

and letting $\delta \to 0$ we obtain

$$\kappa\int_{Q_r(t_0,x_0)\cap\{u>\epsilon\}} u^{-p} \;\leq\; C_4\, r^{n+2-\frac{2p}{1-\gamma}} - \frac{1}{\epsilon}\int_{Q_r(t_0,x_0)\cap\{0<u\leq\epsilon\}} u\,\partial_t w$$

$$-\int_{Q_r(t_0,x_0)\cap\{u>\epsilon\}} \partial_t w \;\leq\; C_4\, r^{n+2-\frac{2p}{1-\gamma}}$$

$$-\frac{1}{\epsilon}\int_{Q_r(t_0,x_0)} \frac{1-\gamma-p}{2-\gamma-p}\,\partial_t \min(\epsilon^{2-\gamma-p}, u^{2-\gamma-p})$$

$$+\int_{B_r(x_0)} \max(\epsilon^{1-\gamma-p}, w)(t_0-r^2) \;-\; \int_{B_r(x_0)} \max(\epsilon^{1-\gamma-p}, w)(t_0+r^2).$$

Integrating by parts in time and letting $\epsilon \to 0$ the lemma is proved. □

From Lemma 2 of [Ph] as well as Lemma 5.4 we obtain now that $u^\gamma|\partial_t u| \leq C u^{2\gamma}$ is locally contained in L^1 provided that $-2\gamma < \frac{1-\gamma}{2}$, which is the case since we assumed

$\gamma \in (-\frac{1}{3}, 0)$. We use this to prove that u is a variational solution in the sense of Definition 3.1: Since the variational formulation does not contain any discontinuous or singular term it is sufficient to prove that $u_\epsilon \to u$ strongly in $L^2((0,\infty); H^{1,2}(\mathbf{R}^n))$ and that $\partial_t u_\epsilon \to \partial_t u$ strongly in $L^2((0,\infty)\times\mathbf{R}^n)$. To this end we take u_ϵ and $\max(u,\delta)$ as test functions for the weak equation of u_ϵ and u, respectively, to obtain (after $\delta \to 0$) the convergence of the L^2-norm of ∇u_ϵ to that of ∇u, and we take $\partial_t u_\epsilon$ and $\partial_t u$ as test functions for the weak equation of u_ϵ and u, respectively, to obtain

$$\int_0^\infty \int_{\mathbf{R}^n} |\partial_t u_\epsilon|^2 \, - \, \frac{1}{2} \int_{\mathbf{R}^n} |\nabla u_\epsilon^0|^2 \, = \, \int_{\mathbf{R}^n} F_\epsilon(u_\epsilon^0)$$

$$\to \int_{\mathbf{R}^n} F_0(u^0) \, = \, \int_0^\infty \int_{\mathbf{R}^n} |\partial_t u|^2 \, - \, \frac{1}{2} \int_{\mathbf{R}^n} |\nabla u^0|^2$$

as the subsequence $\epsilon \to 0$; here $F_\epsilon(z) := \int_0^z f_\epsilon(s)\, ds$ and $F_0(z) := \int_0^z \max(s,0)^\gamma \, ds$. Since $\nabla_{t,x} u_\epsilon \rightharpoonup \nabla_{t,x} u$ weakly in $L^2((0,\infty)\times\mathbf{R}^n)$ it follows that $\nabla_{t,x} u_\epsilon \to \nabla_{t,x} u$ strongly in $L^2((0,\infty)\times\mathbf{R}^n)$ and that u is a variational solution in the sense of Definition 3.1.

As before, the estimates required in the assumptions of Theorem 4.1 follow from the regularity in time and space and from the fact that $u \in L^\infty((0,\infty)\times\mathbf{R}^n)$. Thus, supposing u to be a solution of (5.2) with respect to $\gamma \in (-\frac{1}{3}, 0)$ constructed by Phillips' regularization the blow-up limits are characterized as nontrivial self-similar variational solutions in $\mathbf{R}^{n+1} \cap \{t \le 0\}$ as follows.

For $(T, x_0) \in \partial\{u > 0\} \cap ((0,\infty)\times\mathbf{R}^n)$ and $0 < \rho_k \to 0$ as $k \to \infty$ we have for any open $D \subset\subset \mathbf{R}^{n+1}$ the following: the blow-up sequence

$$u_k(t,x) = \frac{u(T + \rho_k^2 t, x_0 + \rho_k x)}{\rho_k{}^{\frac{2}{1-\gamma}}} \tag{5.3}$$

is bounded in $C^{0,\frac{1}{2}}(\bar{D})$ and any pointwise limit u_0 with respect to a subsequence is a nontrivial nonnegative variational solution of (5.2) which is homogeneous of degree $\frac{2}{1-\gamma}$ on paths $\theta \mapsto (\theta^2 t, \theta x)$ for positive θ and $(t,x) \in \mathbf{R}^{n+1} \cap \{t \le 0\}$ and which preserves the regularity estimates in time and space.

That the blow-up limit is again a variational solution can be inferred from the fact that $u_k \to u_0$ strongly in $H^{1,2}(D)$, which can in turn be proved as before.

We proceed to the heat equation with Bernoulli-type condition on the free boundary which has been suggested in [CaVa] as a model for flame propagation. Here we consider variational solutions of

$$\begin{aligned} &u \ge 0\,,\ \partial_t u - \Delta u = 0 \text{ in } ((0,\infty)\times\mathbf{R}^n) \cap \{u > 0\}, \text{ and} \\ &\qquad |\nabla u| = 1 \text{ on } ((0,\infty)\times\mathbf{R}^n) \cap \partial\{u > 0\}. \end{aligned} \tag{5.4}$$

Since in general singularities appear in finite time there exists no global strong solution of (5.4). In [CaVa] Caffarelli and Vazquez introduce a notion of a weak solution for which they show existence in the case of an advancing flame ($\partial_t u \le 0$).

Here we use the notion of nondegenerate variational solutions in the sense of Definition 3.1 with respect to $p = \lambda_- = 0$. In this context, too, the question of existence with respect to general initial data remains unanswered: that a limit u of the solutions u_ϵ of the regularized equation in [CaVa] is a variational solution is in fact equivalent to the condition that no energy loss occurs in the limit, i.e.,

$$\limsup_{\epsilon\to 0} \int \left((\partial_t u_\epsilon)^2 \, + \, \chi_{\{u_\epsilon > 0\}}\right) \, \le \, \int \left((\partial_t u)^2 \, + \, \chi_{\{u > 0\}}\right)$$

(where $\epsilon \to 0$ is the subsequence with respect to u), which is very much like the situation in mean curvature flow or the Mullins–Sekerka equation—compare to [LuSt]. Notice that this may actually be used to construct numerical solutions since a sudden drop of the energy of the approximate solutions (as ϵ and the parameter h of the Galerkin approximation become small) can be registered numerically. Here we are going to show that no energy loss happens when passing to the blow-up limit of nondegenerate variational solutions, which means that the class of nondegenerate variational solutions is closed with respect to the blow-up process.

Concerning the subsequent nondegeneracy condition we like to point out that in the stationary case a nondegenerate variational solution can always be found, namely, a minimizer of the energy (compare to Lemma 3.4 of [AC] and Theorem 4.1 of [BCN]). In the nonvariational case, especially in the time-dependent case, however, the absence of a forcing term of order -1 makes for a severe difficulty: obviously any constant $> \epsilon$ solves the regularized equation in [CaVa] with respect to suitable initial data and violates the nondegeneracy of u_ϵ. Therefore a nondegeneracy assumption arises in a natural way when introducing weak solutions of the problem; compare to Definition 5.1 in [AC] for the stationary case.

THEOREM 5.1. *Let $u \in H^{1,\frac{1}{2}}((0,\infty)\times\mathbf{R}^n)$ be a variational solution of* (5.4)*, i.e., a nonnegative solution in the sense of Definition* 3.1 *with respect to $p = \lambda_- = 0$ and $\lambda_+ = 1$, and suppose that $(T, x_0) \in ((0,\infty)\times\mathbf{R}^n) \cap \partial\{u > 0\}$ and that $0 < \rho_k \to 0$ as $k \to \infty$. Then for any open $D \subset\subset \mathbf{R}^{n+1}$ the sequence*

$$u_k(t,x) = \frac{u(T+\rho_k^2 t, x_0 + \rho_k x)}{\rho_k}$$

is bounded in $C^{0,\frac{1}{2}}(\bar{D})$ and any pointwise limit u_0 with respect to a subsequence is a function $\in H^{1,\frac{1}{2}}_{\mathrm{loc}}((-\infty,\infty)\times\mathbf{R}^n)$ homogeneous of degree 1 on paths $\theta \mapsto (\theta^2 t, \theta x)$ for positive θ and $(t,x) \in \mathbf{R}^{n+1} \cap \{t \le 0\}$.

Moreover, if u satisfies the nondegeneracy condition that for any open $D \subset\subset (0,\infty)\times\mathbf{R}^n$ there exists $c > 0$ and $\delta > 0$ such that for any $Q_r(t,x) \subset D$ satisfying $r \le \delta$ the implication

$$\sup_{Q_r(t,x)} u \le c\, r \Rightarrow u = 0 \text{ in } Q_{\frac{r}{2}}(t,x) \tag{5.5}$$

holds, then any pointwise limit u_0 with respect to a subsequence is in $\mathbf{R}^{n+1} \cap \{t < 0\}$ a nondegenerate variational solution of (5.4).

Proof. The fact that $u \in H^{1,\frac{1}{2}}_{\mathrm{loc}}((0,\infty)\times\mathbf{R}^n)$ along with the sublinear growth in space imply that the growth estimates of Theorems 3.1 and 4.1 are satisfied in the past of any free boundary point $(T, x_0) \in (0,\infty)\times\mathbf{R}^n$. The point therefore is to show that u_0 is again a variational solution. To this end, we first prove that $\chi_{\{u_k>0\}} \to \chi_{\{u_0>0\}}$ a.e. in $\mathbf{R}^{n+1}$ as (having passed to an appropriate subsequence) $k \to \infty$.

First observe that $Q_{r_1}(t_1,x_1) \subset ((-\infty,+\infty)\times\mathbf{R}^n) \cap \{u_0 > 0\}$ and $Q_{r_2}(t_2,x_2) \subset\subset ((-\infty,+\infty)\times\mathbf{R}^n)\cap\{u_0 = 0\}$ imply by the uniform regularity and nondegeneracy of the sequence u_k that $u_k \ge c_1 > 0$ in $Q_{\frac{r_1}{2}}(t_1,x_1)$ and $u_k = 0$ in $Q_{\frac{r_2}{2}}(t_2,x_2)$ for large k. Thus $\chi_{\{u_k>0\}} = 1$ in $Q_{\frac{r_1}{2}}(t_1,x_1)$ and $\chi_{\{u_k>0\}} = 0$ in $Q_{\frac{r_2}{2}}(t_2,x_2)$ for large k and it is therefore sufficient to prove $\mathcal{L}^{n+1}(\partial\{u_0 > 0\}) = 0$ in order to verify the a.e. convergence of $\chi_{\{u_k>0\}}$. But this follows from the positive Lebesgue density of the set

$\{u_0 > 0\}$ in free boundary points. For any $D \subset\subset (-\infty,\infty)\times \mathbf{R}^n$, $(t,x) \in D \cap \partial\{u_0 > 0\}$, and $r \le \delta_1$, we have

$$\frac{\mathcal{L}^{n+1}(\{u_0>0\}\cap Q_r(t,x))}{\mathcal{L}^{n+1}(Q_r)} \ge c_2 > 0 : \tag{5.6}$$

since u_0 is again nondegenerate, there is $(t_3,x_3) \in Q_r(t,x)$ such that $u_0(t_3,x_3) \ge c\, r$. Therefore the fact that $u_0 \in \mathbf{H}^{1,\frac{1}{2}}(B_{2\delta}(D))$ implies that $u_0 > 0$ in $Q_{c_3 r}(t_3,x_3)$, where $c_3 \in (0,1)$ is a constant depending only on c as well as on the Hölder norm of u_0. Now, since $\mathcal{L}^{n+1}$-a.a. points in $(-\infty,+\infty)\times\mathbf{R}^n$ are Lebesgue-points with respect to $\chi_{\partial\{u_0>0\}}$ we know that

$$\frac{\mathcal{L}^{n+1}(\partial\{u_0>0\}\cap E_r(t,x))}{\mathcal{L}^{n+1}(E_r)} \to 1 \ \text{ as } r\to 0$$

for $\mathcal{L}^{n+1}$-a.a. points $(t,x) \in ((-\infty,+\infty)\times\mathbf{R}^n)\cap\partial\{u_0>0\}$ and $E_r(t,x) := (t-r,t+r)\times B_r(x)$. Considering such a point (t,x) and a subsequence $r\to 0$ such that $\frac{1}{2r} \in \mathbf{N}$ we may therefore decompose (up to a set of vanishing $\mathcal{L}^{n+1}$-measure) $E_r(t,x)$ into $\frac{1}{r}$ parabolic cylinders $Q_r(t_i,x_i)$ $(1 \le i \le \frac{1}{r})$ and conclude that $Q_{\frac{r}{2}}(t_i,x_i)\cap\partial\{u_0>0\} \neq \emptyset$ for at least $\frac{1}{2r}$ cylinders $Q_{\frac{r}{2}}(t_i,x_i)$ if r has been chosen small enough. With respect to those cylinders we may apply (5.6) to obtain

$$\frac{\mathcal{L}^{n+1}(\{u_0>0\}\cap Q_r(t_i,x_i))}{\mathcal{L}^{n+1}(Q_r)} \ge c(n)\, c_2 > 0.$$

Taking the sum with respect to the $\frac{1}{2r}$ cylinders we see that

$$\frac{\mathcal{L}^{n+1}(\{u_0>0\}\cap E_r(t,x))}{\mathcal{L}^{n+1}(E_r)} \ge \frac{1}{2}\, c(n)\, c_2 > 0$$

if r has been chosen small enough, a contradiction to

$$\frac{\mathcal{L}^{n+1}(\partial\{u_0>0\}\cap E_r(t,x))}{\mathcal{L}^{n+1}(E_r)} \to 1.$$

Thus we obtain a contradiction for $\mathcal{L}^{n+1}$-a.e. point $(t,x) \in ((-\infty,+\infty)\times\mathbf{R}^n)\cap\partial\{u_0>0\}$, proving that $\mathcal{L}^{n+1}(\partial\{u_0>0\}\cap((-\infty,+\infty)\times\mathbf{R}^n)) = 0$. Note that in the just-finished proof of the convergence $\chi_{\{u_k>0\}} \to \chi_{\{u_0>0\}}$ a.e. we did not make use of u_k being a blow-up sequence. The same remains true regarding the strong convergence of u_k to u_0 in $L^2_{\mathrm{loc}}((-\infty,\infty); H^{1,2}_{\mathrm{loc}}(\mathbf{R}^n))$: since u_k solves for given t_1,t_2 and large k the heat

equation in the open set $((t_1,t_2)\times\mathbf{R}^n)\cap\{u_k>0\}$, multiplication by $\eta\max(u_k-\delta,0)$ yields for $\delta>0$ and $\eta\in C_0^\infty(\mathbf{R}^n)$

$$\int_{t_1}^{t_2}\int_{\mathbf{R}^n} |\nabla\max(u_k-\delta,0)|^2\,\eta \;=\; \int_{t_1}^{t_2}\int_{\mathbf{R}^n} (-\partial_t u_k\ \max(u_k-\delta,0)\,\eta$$

$$-\max(u_k-\delta,0)\,\nabla u_k\cdot\nabla\eta \;=\; \int_{\mathbf{R}^n}\eta\left(\frac{1}{2}\max(u_k-\delta,0)^2(t_1)-\frac{1}{2}\max(u_k-\delta,0)^2(t_2)\right)$$

$$-\int_{t_1}^{t_2}\int_{\mathbf{R}^n}\max(u_k-\delta,0)\,\nabla u_k\cdot\nabla\eta\,,$$

and the analogous identity holds for u_0. Letting in both identities first $\delta \searrow 0$ and then the subsequence $k \to \infty$ we obtain

$$\int_{t_1}^{t_2} \int_{\mathbf{R}^n} |\nabla u_k|^2 \eta \;\to\; \int_{t_1}^{t_2} \int_{\mathbf{R}^n} |\nabla u_0|^2 \eta,$$

which, together with the weak convergence of u_k to u_0 in $L^2((t_1,t_2); H^{1,2}(B_R(0)))$ for $R \in (0,\infty)$, proves the strong convergence of u_k to u_0 in $L^2_{\mathrm{loc}}(H^{1,2}_{\mathrm{loc}})$. Next we use the monotonicity Theorem 3.1 to prove the strong convergence of $\partial_t u_k$ to $\partial_t u_0$ in $L^2_{\mathrm{loc}}(\mathbf{R}^{n+1} \cap \{t<0\})$ which will, incidentally, complete the proof of our theorem. For some $R, S \in (0,\infty)$ let E be any subset of $(-4R^2, -\frac{9}{4}R^2) \times B_S(0)$ such that E is $\mathcal{L}^{n+1}$-measurable. Then Theorem 3.1 implies that

$$\int_E (2t\,\partial_t u_k)^2 \;=\; \int_E (\nabla u_k \cdot x \;+\; 2t\partial_t u_k \;-\; u_k)^2$$

$$-\; \int_E (\nabla u_k \cdot x \;-\; u_k)^2 \;-\; \int_E 4t\partial_t u_k\,(\nabla u_k \cdot x \;-\; u_k)$$

$$\leq\; 2\,\|\nabla u_k \cdot x \;-\; u_k\|^2_{L^2(E)} \;+\; \frac{1}{2}\int_E (2t\,\partial_t u_k)^2$$

$$+\, C(n,R,S) \int_R^{\frac{3}{2}R} r^{-3} \int_{T_r^-(0)} \frac{1}{-t} (\nabla u_k \cdot x \;+\; 2t\partial_t u_k \;-\; u_k)^2\, G_{(0,0)}\, dr$$

$$\leq\; \frac{81\,R^4\,\epsilon}{32} \;+\; \frac{1}{2}\int_E (2t\,\partial_t u_k)^2$$

$$+\, C(n,R,S) \int_R^{\frac{3}{2}R} r^{-3} \int_{T_r^-(0)} \frac{1}{-t} (\nabla u_k \cdot x \;+\; 2t\partial_t u_k \;-\; u_k)^2\, G_{(0,0)}\, dr$$

provided that $\mathcal{L}^{n+1}(E) \leq \delta$. Since the third term on the right-hand side goes to 0 as $k \to \infty$, we obtain for any positive ϵ positive constants δ and $k_0 < \infty$ such that $\int_E (\partial_t u_k)^2 \;\leq\; \frac{\epsilon}{2}$ provided that $\mathcal{L}^{n+1}(E) \leq \delta$ and $k \geq k_0$. Last, we infer from the a.e. convergence $\partial_t u_k$ to $\partial_t u_0$, which can be proved exactly as the a.e. convergence of $\chi_{\{u_k>0\}}$ to $\chi_{\{u_0>0\}}$ (using the fact that $\mathcal{L}^{n+1}(\partial\{u_0 > 0\}) = 0$ as well as the nondegeneracy of u_k), that $\partial_t u_k \to \partial_t u_0$ in $L^2_{\mathrm{loc}}(\mathbf{R}^{n+1} \cap \{t<0\})$. □

Let us finally estimate the Hausdorff dimension of the free boundary of variational solutions of (5.2) and (5.4).

LEMMA 5.5. *Let u be either a variational solution of* (5.4) *satisfying the nondegeneracy* (5.5) *as well as the regularity $u \in \mathbf{H}^{1,\frac{1}{2}}((0,\infty)\times\mathbf{R}^n)$, or let u be a solution of* (5.2) *with respect to $\gamma \in (-\frac{1}{3}, 0)$ derived by Phillips' approximation. Then for any $(T, x_0) \in ((0,+\infty)\times\mathbf{R}^n) \cap \partial\{u > 0\}$, any sequence $0 < \rho_k \to 0$ and any open $D \subset\subset \mathbf{R}^{n+1}$ the blow-up sequence*

$$u_k(t,x) := \frac{u(T + {\rho_k}^2 t, x_0 + \rho_k x)}{{\rho_k}^{\frac{2}{1-\gamma}}}$$

(here $\gamma := -1$ in the case of a solution of (5.4)) is bounded in $C^{0,\frac{1}{2}}(\bar{D})$ and any pointwise limit u_0 with respect to a subsequence is a nontrivial variational solution in $\mathbf{R}^{n+1} \cap \{t < 0\}$ again satisfying the respective regularity and nondegeneracy and being homogeneous of degree $\frac{2}{1-\gamma}$ on paths $\theta \mapsto (\theta^2 t, \theta x)$ for positive θ and $(t,x) \in \mathbf{R}^{n+1} \cap \{t \le 0\}$. Furthermore for every compact set $K \subset \mathbf{R}^{n+1}$ and every open set $U \supset K \cap \partial\{u_0 > 0\}$ there exists $k_0 < \infty$ such that $\partial\{u_k > 0\} \cap K \subset U$ for $k \ge k_0$.

Note that the last statement is stronger than the one direction in the definition of the convergence in Hausdorff distance in that it holds for *any* open set U.

Proof. Except for the last statement everything follows directly from what was shown before. Now suppose that there exists $(t_k, x_k) \in \partial\{u_k > 0\} \cap (K - U)$. Then $(t_k, x_k) \to (\bar{t}, \bar{x}) \in (K-U) \cap \{u_0 = 0\}$ as a subsequence $k \to \infty$. Assuming now that $(\bar{t}, \bar{x})$ is an inner point of $\{u_0 = 0\}$ the uniform convergence of u_k to u_0 as well as the nondegeneracy of u_k yields a contradiction to the assumption that (t_k, x_k) is a free boundary point of u_k. □

Let us define with respect to the parabolic metric Hausdorff measures

$$\mathcal{H}_m^{\delta,par}(A) := \inf\left\{\sum_{j=1}^{\infty}(\text{pardiam } S_j)^m : A \subset \bigcup_{j=1}^{\infty} S_j \text{ , pardiam } S_j \le \delta\right\}$$

$$\text{and } \mathcal{H}_m^{par}(A) := \lim_{\delta \to 0} \mathcal{H}_m^{\delta,par}(A),$$

where $\delta \in (0, +\infty]$ and pardiam is the diameter in $\mathbf{R}^{n+1}$ with respect to the parabolic metric $\text{dist}((t,y),(s,x)) := \max(|x-y|, |s-t|^{\frac{1}{2}})$. We need three elementary properties of these parabolic Hausdorff measures of which the first as well as the third can be proved exactly as in Lemmas 11.2 and 11.5 of [Giu] and a proof of the second property (closely following Proposition 11.3 of [Giu]) is contained in the appendix:

(5.7) For every $A \subset \mathbf{R}^{n+1}$ the equivalence
$\mathcal{H}_m^{par}(A) = 0$ if and only if $\mathcal{H}_m^{\infty,par}(A) = 0$ holds.

(5.8) For every $n \ge 1$, $m > n$ and $A \subset \mathbf{R}^{n+1}$
$\limsup_{r \to 0} \frac{\mathcal{H}_m^{\infty,par}(A \cap \overline{Q_r^-(t,x)})}{r^m} \ge 1$ for $\mathcal{H}_m^{par}$-a.a. points $(t,x) \in A$.

(5.9) In the situation of Lemma 5.5 we have
$\mathcal{H}_m^{\infty,par}(K \cap \partial\{u_0 > 0\}) \ge \limsup_{k \to \infty} \mathcal{H}_m^{\infty,par}(K \cap \partial\{u_k > 0\})$,

where the limit superior is taken with respect to the subsequence chosen in Lemma 5.5.

Having listed the necessary tools we may now use the dimension reduction procedure. Suppose that $m > n+1$ and $\mathcal{H}_m^{par}(\partial\{u > 0\}) > 0$. Then in $\mathcal{H}_m^{par}$-a.a. points of $\partial\{u > 0\}$ we can use (5.8) as well as (5.9) to obtain a blow-up limit u_0 with the properties mentioned in Lemma 5.5, satisfying $\mathcal{H}_m^{\infty,par}(\partial\{u_0 > 0\} \cap \{t \le 0\}) > 0$. Therefore we find (by (5.7)) a point $(\bar{t}, \bar{x}) \in (\partial\{u_0 > 0\} \cap \{t \le 0\}) - \{(0,0)\}$ in which the density in (5.8) is estimated from below. Now any blow-up limit u_{00} with respect to $(\bar{t}, \bar{x})$ (and with respect to a subsequence such that the limit superior in (5.8) becomes a limit) again satisfies the properties of Lemma 5.5; in addition, we get from the homogeneity of u_0 as in Lemma 3.1 of [We] the following.

If $\bar{t} \neq 0$, then u_{00} has to be constant in time direction and we proceed with $\bar{u} := u_{00}|_{\{t=0\}}$, which in this case is a stationary variational solution satisfying $\mathcal{H}^{m-2}(\partial\{\bar{u} > 0\}) > 0$ as well as the properties of Lemma 5.5.

If $\bar{t} = 0$, then, applying a rotation in space, we may assume that u_{00} is constant in direction of the nth unit vector and proceed with $\bar{u} := u_{00}|_{\{x_n=0\}}$, which is then a variational solution in $(-\infty, +\infty) \times \mathbf{R}^{n-1}$ satisfying $\mathcal{H}^{par}_{m-1}(\partial\{\bar{u} > 0\} \cap \{t \leq 0\}) > 0$ as well as the properties of Lemma 5.5.

Repeating this n times either we obtain a stationary variational solution $u^\star : \mathbf{R} \to \mathbf{R}$ which is nontrivial, homogeneous of degree $\frac{2}{1-\gamma}$, has 0 as free boundary point, and satisfies $\mathcal{H}^{m-n-1}(\partial\{u^\star > 0\}) > 0$, or we obtain a space-independent variational solution $u_\star : (-\infty, 0) \to \mathbf{R}$ which is nontrivial, homogeneous of degree $\frac{1}{1-\gamma}$, has 0 as a free boundary point, and satisfies $\mathcal{H}^{\frac{m-n}{2}}(\partial\{u_\star > 0\} \cap \{t \leq 0\}) > 0$. Since both yield a contradiction if $m > n+1$ we proved the following theorem.

THEOREM 5.2. *Let u be either a variational solution of* (5.4) *in the sense of Definition* 3.1 *satisfying the nondegeneracy* (5.5) *as well as the regularity $u \in \mathbf{H}^{1,\frac{1}{2}}((0,\infty) \times \mathbf{R}^n)$ or a solution of* (5.2) *with respect to $\gamma \in (-\frac{1}{3}, 0)$ derived by Phillips' approximation. Then the parabolic Hausdorff dimension of $\partial\{u > 0\}$ does not exceed $n+1$.*

Considering the stationary solution $\max(x_n, 0)$ of (5.4) in $(0,\infty) \times \mathbf{R}^n$ it is obvious that the parabolic dimension $n+1$ in the estimate cannot be improved.

6. Appendix.

LEMMA 6.1. *If $\ell \geq 1$, $k > \ell$, and A is a subset of $\mathbf{R}^{\ell+1}$, then for $\mathcal{H}^{par}_k$-a.a. points $(t,x) \in A$ we have*

$$\limsup_{r \to 0} \frac{\mathcal{H}^{\infty,par}_k(A \cap \overline{Q^-_r(t,x)})}{r^k} \geq 1$$

(where $Q_r(t,x)$ is an $\ell+1$-dimensional cylinder).

Proof. Define $\zeta(S) := (\text{pardiam}(S))^k$ and, for positive τ, δ, and ϵ,

$$B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon) := \{(t,x) \in A : \mathcal{H}^{\delta,par}_k(S^{t-} \cap A) \leq \tau\zeta(S)$$

for all sets $S \subset \mathbf{R}^{\ell+1}$ such that $S \ni (t,x)$ and $\text{pardiam}(S) < \epsilon\}$ (here $S^{t-} := \{(s,y) \in S : s \leq t\}$). Now suppose $S \subset \mathbf{R}^{\ell+1}$ to satisfy $\text{pardiam}(S) < \epsilon$. Defining $t := \sup\{\sigma : \{s = \sigma\} \cap S \cap B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon) \neq \emptyset\}$ assume first that $\{\sigma = t\} \cap S \cap B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon) = \emptyset$, in which case we obtain for $\theta \in (0, \min(\epsilon, \delta))$

$$\mathcal{H}^{\delta,par}_k(S \cap B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon)) \leq \mathcal{H}^{\delta,par}_k(S^{(t-\theta^2)-} \cap B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon))$$

$$+ \mathcal{H}^{\delta,par}_k((t-\theta^2, t) \times B_\epsilon(x)) \leq \tau\zeta(S) + C(\ell)\,\theta^k \left(\frac{\epsilon}{\theta}\right)^\ell = \tau\zeta(S) + o(1)$$

(as $\theta \to 0$) by the assumption $k - \ell > 0$.

In case there exists $(t,x) \in S \cap B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon)$ we obtain by the definition of $B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon)$ that

$$\mathcal{H}^{\delta,par}_k(S \cap B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon)) \leq \tau\zeta(S).$$

The definition of $\mathcal{H}^{\epsilon,par}_k$ as well as the subadditivity of $\mathcal{H}^{\delta,par}_k$ therefore imply

$$\mathcal{H}^{\delta,par}_k(B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon)) \leq \tau\mathcal{H}^{\epsilon,par}_k(B(\mathcal{H}^{\delta,par}_k, \tau, \epsilon)).$$

In particular, if $\delta < 1$,

$$\mathcal{H}_k^{\delta,par}(B(\mathcal{H}_k^{\delta,par}, 1-\delta, \delta)) \le (1-\delta)\, \mathcal{H}_k^{\delta,par}(B(\mathcal{H}_k^{\delta,par}, 1-\delta, \delta)) \ < \ \infty.$$

Hence $\mathcal{H}_k^{\delta,par}(B(\mathcal{H}_k^{\delta,par}, 1-\delta, \delta)) = 0$ and, by (5.7), $\mathcal{H}_k^{par}(B(\mathcal{H}_k^{\delta,par}, 1-\delta, \delta)) = 0$.

Now consider the set

$$C := \left\{ (t,x) \in A \ : \ \inf_{\epsilon>0} \sup \left\{ \frac{\mathcal{H}_k^{\infty,par}(A \cap S^{t-})}{\zeta(S)} \ : \ S \subset \mathbf{R}^{\ell+1} \text{ such that } \right.\right.$$

$$\left.\left. S \ni (t,x) \text{and pardiam}(S) < \epsilon \right\} \ < 1 \right\}$$

$$= \left\{ (t,x) \in A \ : \ \inf_{n \in \mathbf{N}} \sup \left\{ \frac{\mathcal{H}_k^{\infty,par}(A \cap S^{t-})}{\zeta(S)} \ : \ S \subset \mathbf{R}^{\ell+1} \text{ such that } \right.\right.$$

$$\left.\left. S \ni (t,x) \text{and pardiam}(S) < \frac{1}{n} \right\} \ < 1 \right\}.$$

If $(t,x) \in C$ we must have for some n

$$\sup \left\{ \frac{\mathcal{H}_k^{\infty,par}(A \cap S^{t-})}{\zeta(S)} \ : \ S \subset \mathbf{R}^{\ell+1} \text{ such that } \right.$$

$$\left. S \ni (t,x) \text{and pardiam}(S) < \frac{1}{n} \right\} \ < 1 - \frac{1}{n}.$$

Hence $(t,x) \in B(\mathcal{H}_k^{\infty,par}, 1-\frac{1}{n}, \frac{1}{n}))$, and since the definition of $\mathcal{H}_k^{\delta,par}$ implies $\mathcal{H}_k^{\delta,par}(\tilde{S} \cap A) = \mathcal{H}_k^{\infty,par}(\tilde{S} \cap A)$ for any $\tilde{S}$ such that pardiam$(\tilde{S}) < \delta$, we obtain $C \subset \bigcup_{n=1}^{\infty} B(\mathcal{H}_k^{\infty,par}, 1-\frac{1}{n}, \frac{1}{n})) \subset \bigcup_{n=1}^{\infty} B(\mathcal{H}_k^{\frac{1}{n},par}, 1-\frac{1}{n}, \frac{1}{n}))$. Consequently $\mathcal{H}_k^{par}(C) = 0$.

Finally, fixing $(t,x) \in A$, considering $S \ni (t,x)$, and setting $d := \text{pardiam}(S)$, we obviously have

$$\frac{\mathcal{H}_k^{\infty,par}(A \cap S^{t-})}{\zeta(S)} \ \le \ \frac{\mathcal{H}_k^{\infty,par}(A \cap \overline{Q_d^-(t,x)})}{d^k}.$$

Hence

$$\sup \left\{ \frac{\mathcal{H}_k^{\infty,par}(A \cap S^{t-})}{\zeta(S)} \ : \ S \subset \mathbf{R}^{\ell+1} \text{ such that } \right.$$

$$\left. S \ni (t,x) \text{ and pardiam}(S) < \delta \right\} \le \ \sup \left\{ \frac{\mathcal{H}_k^{\infty,par}(A \cap \overline{Q_d^-(t,x)})}{d^k} \ : \ d < \delta \right\}.$$

Thus the set $D := \{(t,x) \in A \ : \ \limsup_{r \to 0} \frac{\mathcal{H}_k^{\infty,par}(A \cap \overline{Q_d^-(t,x)})}{d^k} < 1\}$ must be contained in C and it follows that $\mathcal{H}_k^{par}(D) = 0$. □

Acknowledgments. The basic questions leading to this paper were asked during a one-week stay at Hokkaido University with Y. Giga, whom we thank for advice and numerous discussions. A vital literature hint was given by S. Müller. Many people, including S. Angenent, H. J. Choe, S. Luckhaus, and H. Matano, contributed helpful discussions.

REFERENCES

[Al] W. K. Allard, *On the first variation of a varifold.* Ann. of Math., 95 (1972), pp. 417–491.

[AC] H. W. Alt and L. A. Caffarelli, *Existence and regularity for a minimum problem with free boundary*, J. Reine Angew. Math., 325 (1981), pp. 105–144.

[AlPh] H. W. Alt and D. Phillips, *A free boundary problem for semilinear elliptic equations*, J. Reine Angew. Math., 368 (1986), pp. 63–107.

[BCN] H. Berestycki, L. A. Caffarelli, and L. Nirenberg, *Uniform estimates for regularization of free boundary problems*, in Analysis and Partial Differential Equations, Marcel Dekker, New York, 1990, pp. 567–619.

[Ca] L. A. Caffarelli, *The regularity of free boundaries in higher dimensions*, Acta Math., 139 (1978), pp. 567–619.

[CaVa] L. A. Caffarelli and J. L. Vazquez, *A free boundary problem for the heat equation arising in flame propagation*, Trans. Amer. Math. Soc., 347 (1995), pp. 411–441.

[CMM] X.-Y. Chen, H. Matano, and M. Mimura, *Finite-point extinction and continuity of interfaces in a nonlinear diffusion equation with strong absorption*, J. Reine Angew. Math., 459 (1995), pp. 1–36.

[ChWe] H. J. Choe and G. S. Weiss, *A semilinear parabolic equation with free boundary*, preprint, submitted 1997.

[GiKo] Y. Giga and R. V. Kohn, *Asymptotically self-similar blow-up of semilinear heat equations*, Comm. Pure Appl. Math., 38 (1985), pp. 297–319.

[Giu] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Boston, Basel, Stuttgart, 1984.

[FrHe] A. Friedman and M. A. Herrero, *Extinction properties of semilinear heat equations with strong absorption*, J. Math. Anal. Appl., 124 (1987), pp. 530–546.

[Le] H. E. Levine, *Advances in quenching*, in Nonlinear Diffusion Equations and Their Equilibrium States, Progr. Nonlinear Differential Equations Appl. 7, Birkhäuser, Boston, 1992.

[LSU] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, American Mathematical Society, Providence, RI, 1988.

[LuSt] S. Luckhaus and T. Sturzenhecker, *Implicit time discretization for the mean curvature flow equation*, Calc. Var. Partial Differential Equations, 3 (1995), pp. 253–271.

[Ph] D. Phillips, *Existence of solutions of quenching problems*, Appl. Anal., 24 (1987), pp. 253–264.

[St] M. Struwe, *On the evolution of harmonic maps in higher dimensions*, J. Differential Geom., 28 (1988), pp. 485–502.

[We] G. S. Weiss, *Partial regularity for a minimum problem with free boundary*, J. Geom. Anal., to appear.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 645–677

APPROXIMATION OF THE STOKES DIRICHLET PROBLEM IN DOMAINS WITH CYLINDRICAL OUTLETS*

MARIA SPECOVIUS-NEUGEBAUER†

Abstract. Let $\Omega \subset \mathbf{R}^3$ be a domain with J cylindrical outlets to infinity and $u = (v, p)$ be a solution of the Dirichlet problem for the Stokes system with prescibed flux H_j through the jth outlet. Let $\{\Omega_R\}$ be the set of bounded domains defined by cutting each cylindrical outlet at the distance R from its origin. The problem investigated is how u can be approximated by solutions u^R of boundary problems which are defined on the bounded subdomain Ω_R. On the artificial boundary $\partial\Omega_R \backslash \partial\Omega$ a boundary condition $Bu^R = h$ has to be added. By a method similar to the Schwartz' alternating method, the asymptotic behavior (as R tends to infinity) for $u - u^R$ is investigated for different types of boundary conditions on the cut cross sections. The existence of solutions u^R that are regular up to the edges is shown while using a boundary operator usually related to free boundary problems. For exponentially decaying data asymptotically precise estimates are derived for the difference $u - u^R$; these results hold true for inhomogeneous boundary conditions on the lateral surface $\partial\Omega$ and nonvanishing divergence. For div $v = 0$ and homogeneous boundary conditions on $\partial\Omega$ the case of L^2-forces also is examined.

Key words. approximation problems, Stokes system, artificial boundary conditions

AMS subject classifications. 35Q30, 35A35

PII. S0036141097325083

1. Preliminaries.

1.1. Introduction. Many problems in hydromechanics lead to the investigation of partial differential equations in unbounded domains. One example of such situations is the flow of a viscous fluid through a system of channels. The behavior of the velocity field v of the fluid particles and the pressure distribution p are often described by the nonlinear Navier–Stokes system. In this context the theory of the steady linear part, the so-called Stokes system

$$-\Delta v + \nabla p = f', \quad \text{div } v = f_{n+1} \quad \text{in } \Omega, \quad v = g \quad \text{on the boundary } \partial\Omega, \tag{1.1}$$

plays the fundamental role. Here $\Omega \subset \mathbf{R}^n$, $n \geq 2$, is an unbounded domain with several cylindrical outlets (see section 2 for the exact notations). The vector field f' is a given external force; f_{n+1} may be considered as a distribution of sources and sinks. To obtain unique solutions $u = (v, p)$ to the system (1.1), conditions at infinity have to be added. One possibility, which is also physically reasonable, is to prescribe the flux through each outlet; this condition will be used in the following. Although it is obvious that there exist no infinite volumes of liquid, problems in unbounded domains serve as models for practical problems. On the other hand, any sort of computational work can only be done on finite domains. One approach for overcoming the difficulties with unbounded domains is to truncate the domain, which leads to a problem on a bounded part $\Omega_R \subset \Omega$. For domains with cylindrical outlets, the easiest way to define Ω_R is to cut every outlet at the distance R from its origin. If we assume that the original problem is preserved on Ω_R and on $\partial\Omega_R \cap \partial\Omega$, at least an addititional

*Received by the editors July 23, 1997; accepted for publication (in revised form) March 23, 1998; published electronically April 9, 1999.
http://www.siam.org/journals/sima/30-3/32508.html
†Universität Paderborn, 33095 Paderborn, Germany (mariasp@uni-paderborn.de).

boundary condition

$$B^R u^R = h^R \tag{1.2}$$

has to be imposed on $\partial\Omega_R \backslash \partial\Omega$. Such conditions are usually called *artificial boundary conditions* (ABCs), and we refer to the problem on the bounded domain as an *approximation problem*.

During the last 20 years many efforts were made to create ABCs for various sorts of partial differential equations; we refer here to the survey articles [12, 53]. It is clear that the minimal requirements for the choice of ABCs are the following criteria:

(i) The approximation problem possesses a unique solution $u^R = (v^R, p^R)$.

(ii) On Ω_R the solution u^R is sufficiently close to the solution u of the original problem.

Of course, the best ABCs with respect to the second criterion are the so-called exact ABCs; this means $(u - u^R)|_{\Omega_R} = 0$. However, with the exception of some trivial, i.e., 1-dimensional cases, exact ABCs are *nonlocal*, which means the corresponding boundary operator is a pseudodifferential operator and can be derived in a simple form only for some particular geometries. In practical situations these boundary conditions also have to be approximated (see, e.g., [20, 21, 51, 16, 44] and especially the review articles [12, 53] and the papers cited there). Throughout this paper we deal with local ABCs. This means we investigate different types of boundary conditions in differential form on the artificial boundary $\partial\Omega_R \backslash \partial\Omega$, such that the problem on the truncated domain is elliptic in the sense of Agmon–Douglis–Nirenberg. These conditions are available for a very general class of data f and g not necessarily compactly supported. Local ABCs taking into account the asymptotic behavior of the solution to the Stokes system in an exterior 3-dimensional domain were proposed in [15, 8, 35]. For the Stokes problem in domains with outlets to infinity, a systematic study of local ABCs currently does not seem to exist.

Although the asymptotic behavior of $u - u^R$ is interesting by itself, independent of convergence, at least the application to computations requires a frame where the solution u (for the problem in the unbounded domain Ω) exists for appropriate data and is uniquely determined. Roughly speaking, solutions of boundary value problems in unbounded domains are not only determined by the right-hand sides of the differential equation system and the boundary values, conditions at infinity also have to be imposed.

Since Poincaré's inequality holds in domains with cylindrical outlets, it is clear that for $f \in H^{-1}(\Omega)^3$ (the space of all continuous linear functionals on $\overset{\circ}{H}{}^1(\Omega)^3$) there always exists a weak solution of the Stokes system (1.1) for $f_{n+1} = 0$ and $g = 0$. This means there is $v \in \overset{\circ}{H}{}^1(\Omega)^3$ with $(\nabla v, \nabla\varphi)_\Omega = (f, \varphi)_\Omega$ for all $\varphi \in C_0^\infty(\Omega)^3$ with $\operatorname{div} \varphi = 0$. However, using Gauss' theorem in bounded parts of Ω and Hölder's inequality, it is easy to see that this solution has zero flux through every outlet, which of course is physically nonrealistic. With the help of a *flux carrier* it is possible to construct solutions to the Stokes system (1.1) with prescribed flux H_j through each outlet, but with an unbounded Dirichlet integral, provided the total flux vanishes, i.e., $\sum H_j = 0$. This technique was introduced and applied to a very large class of domains by Ladyženskaja and Solonnikov [27].

A different approach was used by Nazarov and Pileckas in [36]. They applied the results for general elliptic systems in cylinders to derive existence, uniqueness, and asymptotic representations for the system (1.1) with exponentially dying data, where the flux carrier is the Poiseuille flow for large x. In this context they developed a

setting in which more general asymptotic conditions (than conditions on fluxes) can be posed at infinity. We use their results for the problem with prescribed fluxes.

To approximate these solutions we define a domain Ω_R by cutting the outlets at some distance from the origin. As already mentioned, it is reasonable to demand that the approximation problem has a unique solution converging to the solution $u = (v, p)$ of (1.1) with prescribed fluxes in some sense. Moreover, the approximating solution should preserve as much as possible from regularity properties of u. For domains with cylindrical outlets this is a problem insofar as the artificial boundary $\partial\Omega_R \backslash \partial\Omega$ and the cutted boundary $\partial\Omega \backslash \partial\Omega_R$ meet each other, usually in an edge. Even if the boundary of the truncated domain inherits the regularity of the boundary $\partial\Omega$ then the only reasonable situation, where the local H^l-regularity of (v, p) is preserved for arbitrary l, is the case $v|_{\partial\Omega} = 0$ and $v^R|_{\partial\Omega_R \backslash \partial\Omega} = 0$. In all other cases there normally appears a jump in the boundary conditions while using a "smooth" cutting of Ω, or edges in the domain Ω_R when doing sharp cuts. Both methods diminish the regularity properties of u^R. We will use the second approach and cut the outlets orthogonal to the axis of the cylinders. Then Ω_R has edges with an opening angle $\pi/2$, but we bend the data of the approximating problem smoothly to 0 before reaching the edge. We will see that it is possible to find approximating solutions which are C^∞ in the neighborhood of the cut, if the boundary condition is chosen properly.

The main results of the present paper are the following: using the method of matched asymptotic expansions, which is similar to the alternating method of Schwartz (see [50, Chapter IV, section 2]), the asymptotic behavior of $u - u^R$ is calculated for different classes of local ABCs on the artificial boundary (section 3). The existence of unique solutions to the approximation problem, together with regularity up to the edges, is proved for the approximation problem with an appropriate boundary condition on the artificial boundary (Theorem 4.2 and 4.4). In section 5 we derive a uniform estimate for solutions to the Stokes problem on the truncated domain. This estimate is the main tool to obtain an asymptotically precise estimate for $u|_{\Omega_R} - u^R$ as $R \to \infty$ in the case of exponentially decaying data (Theorem 6.1): the difference decays exponentially in $H^{l+1}(\Omega_R)^3 \times H^l(\Omega_R)$. Moreover, for L^2-data it can be shown at least that $\|v - v^R; H^2(\Omega_R)^3\| + \|\nabla(p - p^R); L^2(\Omega_R)^3\| \to 0$ as $R \to \infty$ (section 7).

1.2. Notations, characterization of the domain, basic function spaces, auxiliary results. We recall the notations of some basic function spaces. Let $\mathcal{G} \subset \mathbf{R}^3$ be an open set with closure $\overline{\mathcal{G}}$ and boundary $\partial\mathcal{G}$, which we assume at least to be Lipschitz; ν denotes the exterior normal vector on $\partial\mathcal{G}$. $C_0^\infty(\mathcal{G})$ denotes the set of all smooth functions with compact support in $\mathcal{G}$. For $l \in \mathbf{N}$, $C^l(\mathcal{G})$ is the space of l times continuously differentiable functions φ. We use the common multi-index terminology for partial derivatives: Let $\alpha \in \mathbf{N}_0^3$ with $|\alpha| = \sum_{i=1}^3 \alpha_i$, then $\partial_x^\alpha \varphi(x) = \partial^\alpha \varphi(x) = \frac{\partial^{\alpha_1}}{\partial x_1} \cdots \frac{\partial^{\alpha_3}}{\partial x_3} \varphi(x)$. In this context we mention that the notation $\partial\mathcal{G} \in C^l$ means: For each $x_0 \in \partial\mathcal{G}$ there exists a neighborhood $\mathcal{O}(x_0)$ and a C^l-diffeomorphism $\Phi : \mathcal{O} \to \mathcal{O}'$, such that $\Phi(\mathcal{O} \cap \partial\mathcal{G})$ is an open subset in $\mathbf{R}^{n-1}$. We indicate the L^2-scalar product on $\mathcal{G}$ and on $\partial\mathcal{G}$ by $(\,\cdot\,,\,\cdot\,)_\mathcal{G}$ and $(\,\cdot\,,\,\cdot\,)_{\partial\mathcal{G}}$, i.e., $(u, U)_\mathcal{G} = \int_\mathcal{G} u(x)\bar{U}(x)\,\mathrm{d}x$, $(u, U)_{\partial\mathcal{G}} = \int_{\partial\mathcal{G}} u(x)\bar{U}(x)\,\mathrm{d}o$, where $\mathrm{d}o$ is the 2-dimensional surface measure on $\partial\mathcal{G}$. We extend this notation to all measurable functions (vectorfields), where the integral is finite. We arrange the following convention: in calculations, vectors are always columns, otherwise we use the superscript $^\top$.

We use Sobolev spaces generated by L^2-norms, i.e., $H^l(\mathcal{G})$ is the space consisting of all functions $\varphi \in L^2(\mathcal{G})$ with $\nabla^m \varphi \in L^2(\mathcal{G})$; for all $0 \le m \le l$, ∇^m denotes the system of all (distributional) derivatives $\partial^\gamma \varphi$ of order m. As usual, we indicate the

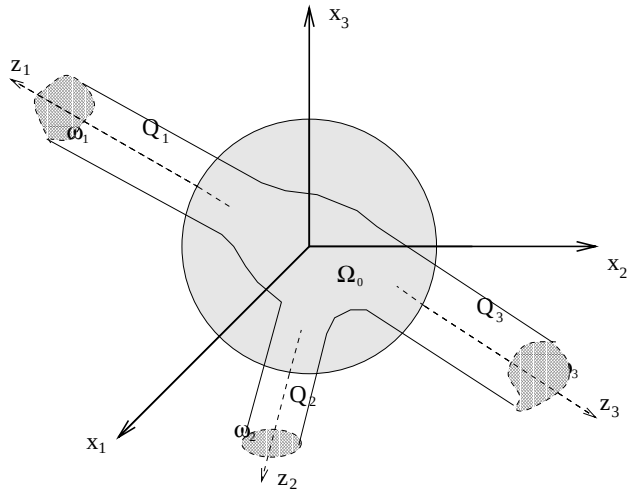


FIG. 1.

closure of $C_0^\infty(\mathcal{G})$ in $H^l(\mathcal{G})$ by $\overset{\circ}{H}{}^l(\Omega)$, the space with zero traces on the boundary. $\varphi \in H^l_{loc}(\mathcal{G})$ means $\varphi|_K \in H^l(K)$ for any bounded open subset of $\mathcal{G}$ with $\overline{K} \subset \mathcal{G}$. If $\mathcal{G}$ is unbounded, we define $H^l_{loc}(\overline{\mathcal{G}})$ in an analogous way ($H^l_{loc}(\overline{\mathcal{G}}) = H^l(\mathcal{G})$ if $\mathcal{G}$ is bounded).

We characterize the underlying domain Ω with J cylindrical outlets to infinity in the following way (see Fig. 1 for $J = 3$). We assume that $\partial\Omega$ is of class C^{l+2} for some $l \in \mathbf{N}$, $\Omega = \Omega_0 \cup Q_1 \cup \cdots \cup Q_J$, where Ω_0 is the intersection of Ω with a ball $B(0, R_0)$ of radius R_0 and $Q_j \cap Q_j = \emptyset$ for $i \neq j$. For each outlet Q_j we introduce a system of local coordinates $(y_j, z_j) \in \omega_j \times [1, \infty)$, where the cross section $\omega_j \subset \mathbf{R}^2$ is a smoothly surrounded bounded domain. Without loss of generality we may assume that $\omega_j \times [2, \infty) \cap \Omega_0 = \emptyset$. When considering a fixed outlet we omit the index j of the local coordinates if no confusion arises. To the cross section ω_j we also assign the cylinder $\Pi_j = \omega_j \times \mathbf{R}$ and the semicylinder $\Pi_j^- = \omega_j \times (-\infty, 0)$.

We recall the main results on existence and uniqueness for the problem (1.1) in the domain Ω. For simplicity assume for the moment that the domain under consideration is a cylinder $\Pi \equiv \omega \times \mathbf{R}$, where $\omega \subset \mathbf{R}^{n-1}$ is a bounded domain, and that the data are smooth and have compact support. We indicate the formal differential operator of the system (1.1) by S, but in order to obtain a formally self-adjoint system, we change div to $-$div , hence $Su = \binom{-\Delta v + \nabla p}{-\operatorname{div} u}$. Performing the complex Fourier transformation $\mathfrak{F}_{z \to \lambda}$ with respect to the variable z along the axis, the problem $S(\nabla_y, \partial_z)u = f$, $v|_{\partial\omega} = g$ turns into a family of problems on the cross section ω of the form

$$S(\nabla_y, i\lambda)U(y) = F(y), \quad y \in \omega; \qquad V(y) = G(y), \quad y \in \partial\omega. \tag{1.3}$$

We denote the associated operator pencil by $\mathfrak{S}(\lambda)$, i.e., $\mathfrak{S}(\lambda)$ is the family of mappings from $\mathcal{D}^l H(\omega) = H^{l+1}(\Omega)^3 \times H^l(\omega)$ to $\mathcal{R}^l(\omega) = H^{l-1}(\Omega)^3 \times H^l(\omega) \times H^{l+1/2}(\omega)^3$ related to the system (1.3). Suppose for $\beta \in \mathbf{R}$ that the line $\mathbf{R} + i\beta$ is free of eigenvalues of the operator pencil $\mathfrak{S}(\lambda)$, which means the problem (1.3) has a unique solution (v_λ, p_λ) for each $\lambda \in \mathbf{R} + i\beta$. Then the inverse Fourier transformation applied to the family (v_λ, p_λ) yields the solution (v, p) for the problem in the cylinder. Parseval's

identity provides

$$\int_{\mathbf{R}} e^{2\beta t}|w(t)|^2\,dt = \int_{\mathbf{R}+i\beta} |\hat{w}(\lambda)|^2\,d\lambda. \tag{1.4}$$

This relation is the motivation to formulate the problem in terms of spaces containing functions with certain exponential weights (for details see, e.g., [33, Chapter 3]).

On Π_j and Ω we define weighted Sobolev spaces in the following way.

DEFINITION 1.1. *Let $l \in \mathbf{N}_0$, $\beta \in \mathbf{R}$, and $\Pi = \omega \times \mathbf{R}$, where ω is a bounded domain in $\mathbf{R}^2$. For $u \in C_0^\infty(\overline{\Pi})$ we set*

$$\|u; W_\beta^l(\Pi)\| := \|ue^{\beta z}; H^l(\Pi)\|.$$

We define $W_\beta^l(\Pi)$ as the closure of $C_0^\infty(\overline{\Pi})$ in this norm.

To extend this definition to the domain Ω with cylindrical outlets we introduce a function $\mathfrak{z} \in C^\infty(\Omega)$ with $\mathfrak{z}(x) = 1$ for $x \in \Omega_0$ and $\mathfrak{z}(x) = z_j$ for $x \in Q_j$ and $z_j \geq 2$.

DEFINITION 1.2. *For l, β as above we define $W_\beta^l(\Omega)$ as the set of all $u \in H_{loc}^l(\Omega)$ such that*

$$\|u; W_\beta^l(\Omega)\| = \|u\, e^{\beta\mathfrak{z}}; H^l(\Omega)\| < \infty.$$

For $l \geq 1$, $\partial\Omega \in C^{l+1}$, we set $W_\beta^{l-1/2}(\partial\Omega) = \{u|_{\partial\Omega} : u \in W_\beta^l(\Omega)\}$ normed by

$$\|u; W_\beta^{l-1/2}(\partial\Omega)\| = \inf \|\tilde{u}; W_\beta^l(\Omega)\|,$$

where the infimum is taken over all $\tilde{u} \in W_\beta^l(\Omega)$ such that $\tilde{u}|_{\partial\Omega} = u$.

Remark. Here we have to observe that if $\Omega = \Pi$ is a cylinder, Definition 1.2 differs from Definition 1.1, when the cylinder Π is regarded as a domain with two outlets.

It is clear that $C_0^\infty(\overline{\Omega})$ is dense in $W_\beta^l(\Omega)$, since this is true for $H^l(\Omega)$ and the weight function $e^{\beta\mathfrak{z}(x)} \geq 1$ for all $x \in \Omega$. We have the obvious embedding $W_\beta^l(\Omega) \hookrightarrow W_{\bar{\beta}}^m(\Omega)$ for $l \geq m$, $\beta \geq \bar{\beta}$ (which is not true for $W_\beta^l(\Pi)$, of course!). Moreover, for $l > m$, $\beta > \bar{\beta}$, this embedding is compact (see the proof of [33, Proposition 4.1.1] and also [52, Lemma 5.4.1]).

We introduce the natural spaces for the data $f = (f', f_{n+1})$, g and the solution $u = (v, p)$ of the system (1.1). For $l \in \mathbf{N}$, let $\partial\Omega \in C^{l+2}$, $\beta \in \mathbf{R}$; we set

$$\begin{aligned} \mathcal{D}_\beta^l W(\Omega) &= W_\beta^{l+1}(\Omega)^3 \times W_\beta^l(\Omega), \\ \mathcal{R}_\beta^l W(\Omega, \partial\Omega) &= W_\beta^{l-1}(\Omega)^3 \times W_\beta^l(\Omega) \times W_\beta^{l+1/2}(\partial\Omega)^3. \end{aligned} \tag{1.5}$$

The mapping

$$\begin{aligned} \mathbb{S}_\beta : \mathcal{D}_\beta^l W(\Omega) &\to \mathcal{R}_\beta^l W(\Omega, \partial\Omega) \\ u &\mapsto (Su, v|_{\partial\Omega}) = (-\Delta v + \nabla p, -\operatorname{div} v, v|_{\partial\Omega}) \end{aligned} \tag{1.6}$$

defines a continuous linear operator. Moreover, the following Green's formula

$$(Su, U)_\Omega - (u, SU)_\Omega = (v, NU)_{\partial\Omega} - (Nu, V)_{\partial\Omega} \tag{1.7}$$

is valid for $u = (v, p) \in \mathcal{D}_\beta^l W(\Omega)$, $U = (V, P) \in \mathcal{D}_{-\beta}^l W(\Omega)$. Here Nu is the Neumann operator, $Nu = (-\nabla v \cdot \nu + p\nu)|_{\partial\Omega}$, ν is the exterior normal vector on $\partial\Omega$. The following results are well known (see [33, Chapter 5.8], [36]).

LEMMA 1.3. *Let $l \in \mathbf{N}$ and $\beta \in \mathbf{R}$. Then the following hold:*

(i) *The mapping* (1.6) *defines a Fredholm operator for all $\beta \in \mathbf{R}$ with the exception of a discrete countable subset $\mathcal{I} \subset \mathbf{R}$.*

(ii) *$0 \in \mathcal{I}$, and $\beta \notin \mathcal{I}$ implies $-\beta \notin \mathcal{I}$. If $\beta \notin \mathcal{I}$, then* $\ker \mathbb{S}_\beta = \operatorname{coker} \mathbb{S}_{-\beta}$ *(in the sense that the necessary conditions arising by Green's formula* (1.7) *are sufficient for the solvability of the system).*

We have $\beta \in \mathcal{I}$ iff the line $\operatorname{Im}\lambda = \beta$ is not free of eigenvalues of the associated elliptic pencils $\mathfrak{S}(\lambda)_j$ in $\Pi_j = \omega_j \times \mathbf{R}$, $j = 1, \dots, J$. This means that there exists λ with $\operatorname{Im}\lambda = \beta$, such that the homogeneous problem (1.3) (i.e., $F = 0$, $G = 0$) has a nontrivial solution on ω_j for at least one j. All eigenvalues give rise to special solutions $u_\lambda^{(k,m)}$ of the homogeneous Stokes system (1.1) in $\Pi_j = \omega_j \times \mathbf{R}$, which are called *exponential solutions of order k*. We have

$$u_\lambda^{(k,m)}(y,z) = e^{i\lambda z} \sum_{q=0}^{k} \frac{1}{q!}(iz)^q \varphi^{(k-q,m)}(y), \tag{1.8}$$

where the vectors $\{\varphi^{(l,m)}(y)\}_{l=1}^k$ form a Jordan chain corresponding to the eigenvalue λ, especially where $\varphi^{(0,m)}$ is a solution to $\mathfrak{S}(\lambda)\varphi^{(0,m)} = 0$ in ω_j. (The index m indicates that there is more than one eigenvector and associated Jordan chain in general.) We recall a further well-known result on the index of the operator $\mathbb{S}_\beta$ and the asymptotic representation of the solutions to (1.1). For this purpose we fix a system of cut-off functions with the properties $\operatorname{supp}\chi_j \subset Q_j$, $\chi_j(y_j, z_j) = \chi_j(z_j) = 1$ for $z_j > 2$ in the local coordinates of Q_j.

LEMMA 1.4 (see [33, Theorem 5.1.4]). *Let $\beta > \gamma$ such that both of the lines $\operatorname{Im}\lambda = \beta$ and $\operatorname{Im}\lambda = \gamma$ are free of eigenvalues of the elliptic pencils $\mathfrak{S}(\lambda)_j$, $j = 1, \dots, J$. Let $\{\lambda_{1,j}, \dots, \lambda_{\mu_j,j}\}$ be the set of eigenvalues of $\mathfrak{S}(\lambda)_j$ with $\beta < \operatorname{Im}\lambda < \gamma$ and $æ(\lambda)$ denote the total multiplicity of the eigenvalue λ. Then the following hold:*

$$\operatorname{Ind}\mathbb{S}_\beta - \operatorname{Ind}\mathbb{S}_\gamma = \sum_{j=1}^{J} \sum_{\mu=1}^{\mu_j} æ(\lambda_{\mu,j}). \tag{1.9}$$

If $(f,g) \in \mathcal{R}_\gamma^l W(\Omega)$, and $u \in \mathcal{D}_\beta^l W(\Omega)$ is a solution to (1.1)*, then u admits the following asymptotic representation as a sum of exponential solutions and $\tilde{u} \in \mathcal{D}_\beta^l W(\Omega)$,*

$$u = \sum_{j=1}^{J} \chi_j \sum_{\mu=1}^{\mu_j} \sum_{m,k} C_{\mu,m,k}(f,g) u_{\lambda_{\mu,j}}^{(k,m)}(y_j, z_j) + \tilde{u}. \tag{1.10}$$

Remark 1.5. Of course, these eigenvalues in general differ from outlet to outlet, but on the real line, $\lambda = 0$ is the only eigenvalue in every cylinder Π_j with corresponding Jordan chain of length 2 (see [36]); we repeat the calculations for the reader's convenience.

Let ω be one of the cross sections ω_j, $y = (y_1, y_2) \in \omega$. The system (1.3) for an eigenvector $\varphi^\top = (\varphi_1, \dots, \varphi_4) \in \mathcal{D}^l H(\omega)$ reads

$$-\Delta_y \varphi_k + \lambda^2 \varphi_k + \partial_k \varphi_4 = 0, \quad k = 1, 2, \tag{1.11}$$

$$-\Delta_y \varphi_3 + \lambda^2 \varphi_3 + i\lambda \varphi_4 = 0, \tag{1.12}$$

$$-\partial_1 \varphi_1 - \partial_2 \varphi_2 - i\lambda \varphi_3 = 0, \tag{1.13}$$

$$\varphi_1|_{\partial\omega} = \varphi_2|_{\partial\omega} = \varphi_3|_{\partial\omega} = 0. \tag{1.14}$$

We multiply the equations (1.11) and (1.12) scalarly in $L^2(\omega)$ by φ_k and φ_3, respectively. In summary, integrating by parts and exploiting (1.13) leads to $0 = \sum_{k=1}^3 \int_\omega (|\nabla_y \varphi_k|^2 + \lambda^2 |\varphi_k|^2)\,\mathrm{d}y$. For any $\lambda \in \mathbf{R}$ this leads, together with the boundary condition (1.14), to $\varphi_k = 0$ for $k = 1, 2, 3$. From (1.12) we obtain $\varphi_4 = 0$ for $\lambda \neq 0$, and $\varphi_4 = const$ for $\lambda = 0$. Hence $\lambda = 0$ is the only eigenvalue of $\mathfrak{S}(\lambda)$ on the real line with corresponding eigenvector $\varphi^{(0)} = (0, 0, 0, 1)^\top$. To calculate the corresponding Jordan chain, we recall the equation for the associated vectors:

$$\mathfrak{S}(0)\varphi^m = -\sum_{l=1}^{m} \frac{1}{l!} \frac{\mathrm{d}^l \mathfrak{S}}{\mathrm{d}\lambda^l}(0)\varphi^{(m-l)}, \quad m = 1, \ldots, \mathfrak{ae} - 1, \tag{1.15}$$

with $\mathfrak{ae}$ as in Lemma 1.4. Differentiating (1.11)–(1.13) with respect to λ, substituting $\lambda = 0$ and $\varphi^{(0)} = (0, 0, 0, 1)^\top$, elementary calculations lead to $\varphi^{(1)} = (0, 0, \frac{i}{2}\Psi, 0)$, where Ψ is a solution to the Dirichlet problem

$$-\Delta_y \Psi = 2 \text{ in } \omega, \qquad \Psi = 0 \text{ on } \partial\omega. \tag{1.16}$$

To see that there exists no associated vector $\varphi^{(2)}$ of order 2, let us assume the contrary. By (1.15), $\varphi^{(2)}$ is a solution to the problem

$$\begin{aligned} -\Delta_y \varphi_k^{(2)} + \frac{\partial}{\partial y_k}\varphi^{(2)} 2_4 &= -\varphi_k^{(0)} = 0, \quad k = 1, 2, \\ -\Delta_y \varphi_3^{(2)} &= -\varphi_3^{(0)} + i\varphi_4^{(1)} = 0, \\ -\frac{\partial}{\partial y_1}\varphi_1^{(2)} - \frac{\partial}{\partial y_2}\varphi_2^{(2)} &= i\varphi_3^{(1)} = -\frac{1}{2}\Psi \text{ in } \omega, \end{aligned} \tag{1.17}$$

$$\varphi_k^{(2)} = 0 \text{ on } \partial\omega, \quad k = 1, 2, 3. \tag{1.18}$$

We integrate (1.17) over ω. From the Gauss theorem and $\Delta\Psi = -2$ we obtain

$$0 = \int_\omega (\frac{\partial}{\partial y_1}\varphi_1^{(2)} + \frac{\partial}{\partial y_2}\varphi_2^{(2)})\,\mathrm{d}y = \frac{1}{2}\int_\omega \Psi\,\mathrm{d}y = -\frac{1}{4}\int \Psi\Delta\Psi\,\mathrm{d}y = -\frac{1}{4}\int_\omega |\nabla\Psi|^2\,\mathrm{d}y > 0,$$

which leads to a contradiction.

The eigenvector and the associated vector give rise to two polynomial (with respect to z) solutions of the homogeneous Stokes system in each cylinder Π_j. These are the *constant pressure solution* $u^{j0}(y, z)^\top = (0, 0, 0, 1)$, and the *Poiseuille flow* $u^{j1}(y, z)^\top = (0, 0, \varpi_j \Psi^{(j)}(y), -2\varpi_j z)$, where $\Psi^{(j)}$ is defined as in (1.16) with $\omega = \omega_j$, and $\varpi_j \in \mathbf{R}$ is chosen in such a way that

$$\int_{\omega_j} \varpi_j \Psi^{(j)}(y)\,\mathrm{d}y = 1, \tag{1.19}$$

i.e., u^{j1} carries the unit flux through the cylinder Π_j.

In the following we fix $l \geq 1$ and $\beta > 0$, $\beta^* > 0$ in such a way that the strips $0 < |\mathrm{Im}\,\lambda| < \beta^*$ are free of eigenvalues of the pencils $\mathfrak{S}(\lambda)_j$ for $j = 1, \ldots, J$, and $0 < \beta < \beta^*$. Since $\mathcal{D}_\beta^l W(\Omega) \subset H^2(\Omega)^3 \times H^1(\Omega)$, it follows from standard methods that the kernel of $\mathbb{S}_\beta$ is trivial, and thus, by Lemma 1.3, the operator $\mathbb{S}_{-\beta}$ is surjective. The index formula (1.9), Lemma 1.3, and Remark 1.5 lead to $\dim\ker\mathbb{S}_{-\beta} + \dim\mathrm{coker}\,\mathbb{S}_\beta = 2\dim\ker\mathbb{S}_{-\beta} = 2J$, hence $\dim\ker\mathbb{S}_{-\beta} = J$. Let $(f, g) \in \mathcal{R}_\beta^l$. We obtain a solution $u \in \mathcal{D}_\beta^l W(\Omega)$ of the system (1.1) iff

$$(f, U)_\Omega - (g, NU)_{\partial\Omega} = 0 \tag{1.20}$$

for all $U \in \ker \mathbb{S}_{-\beta}$. However, since $\mathcal{R}^l_\beta W(\Omega, \partial\Omega) \subset \mathcal{R}^l_{-\beta} W(\Omega, \partial\Omega)$, (1.1) always has a solution in $\mathcal{D}^l_{-\beta} W(\Omega)$ for every $(f, g) \in \mathcal{R}^l_\beta$. By (1.10) and Remark 1.5, we have the following asymptotic representation for u (see [36]):

$$u(x) = \sum_{j=1}^{J} \chi_j(x) \Big((a_j u^{j0}(x) + b_j u^{j1}(x) \Big) + \tilde{u}(x) \tag{1.21}$$

with coefficients $a_j, b_j \in \mathbb{C}$, $\tilde{u} \in \mathcal{D}^l_\beta W(\Omega)$. Here χ_j, $j = 1, \dots, J$, are the same cut-off functions as in Lemma 1.4. In order to simplify the presentation of calculations we adopt the following notations from [36]. We define $4 \times J-$matrices $\mathcal{U}^0$, $\mathcal{U}^1$, with columns $\chi_j u^{jh}$, $h = 0, 1$. For $c \in \mathbb{C}^J$ we have $\mathcal{U}^h \cdot c = \sum_j c_j \chi_j u^{jh}$. In this notation, (1.21) reads

$$u = \mathcal{U}^0 \cdot a + \mathcal{U}^1 \cdot b + \tilde{u}, \qquad a, b \in \mathbb{C}^J. \tag{1.22}$$

Hence we find the preimage $\mathbb{D}^l_{\pm\beta} W(\Omega)$ of $\mathcal{R}^l_\beta W(\Omega, \partial\Omega)$ in $\mathcal{D}^l_{-\beta} W(\Omega)$ as a space with detached asymptotics. Here the index $\pm\beta$ indicates that the space is related both to W^l_β, where the data are taken from, and to $W^l_{-\beta}$, where we look for the solutions. Precisely, we define (see [36])

$$\mathbb{D}^l_{\pm\beta} W(\Omega) = \{u \in \mathcal{D}^l_{-\beta} W(\Omega) : u \text{ fulfills (1.22)}\}.$$

The factor space $\mathbb{D}^l_{\pm\beta} W(\Omega)/\mathcal{D}^l_\beta W(\Omega)$ is isomorphic to $\mathbb{C}^{2J}$ and $\mathbb{D}^l_{\pm\beta} W(\Omega)$ is a Banach space provided with the norm

$$\|u; \mathbb{D}^l_{\pm\beta} W(\Omega)\| = \big(\|\tilde{u}; \mathcal{D}^l_\beta W(\Omega)\|^2 + |a|^2 + |b|^2 \big)^{1/2},$$

where $|\,.\,|$ means the Euclidean norm in $\mathbb{C}^J$. We denote the projections of u on the coefficients (a, b) in (1.22) by πu, i.e., $\pi u = (a, b) =: (\pi_0 u, \pi_1 u)$. By definition, the operator

$$\mathbb{S} : \mathbb{D}^l_{\pm\beta} W(\Omega) \ni u \mapsto (Su, v|_{\partial\Omega}) \in \mathcal{R}^l_\beta W(\Omega)$$

is surjective and $\ker \mathbb{S} = \ker \mathbb{S}_{-\beta}$; moreover, due to Lemma 1.3, $\dim \ker \mathbb{S} = J$. This means in particular that the coefficients a_j, b_j are not uniquely determined by the data. Namely, to the solution we may add $\eta \in \ker \mathbb{S}$, where η, of course, also has the form (1.22). Thus, to obtain unique solutions of the Stokes system in $\mathbb{D}^l_{\pm\beta} W(\Omega)$, additional asymptotic conditions have to be prescribed of the form

$$Bu = \mathbb{B} \cdot (\pi u) = \mathbb{B} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = H \in \mathbb{C}^J, \tag{1.23}$$

where $\mathbb{B}$ is a suitable $J \times 2J$ matrix and $\binom{a}{b}$ is the column of coefficients in the asymptotic representation (1.21). It is self-evident that only special matrices $\mathbb{B}$ are reasonable, i.e., lead to Fredholm properties of the operator

$$(\mathbb{S}, B) : \mathbb{D}^l_{\pm\beta} W(\Omega) \to \mathcal{R}^l_\beta W(\Omega) \times \mathbb{C}^J.$$

We restrict ourselves in the following to the case $\mathbb{B} = (\mathbb{O}, \mathbb{I})$, where $\mathbb{I}$ is the $J \times J$ unit-matrix, and $\mathbb{O}$ is the matrix with all entries 0. This means we prescribe $b = H$. In the case $g = 0$, $f_4 = 0$ this is the problem with prescribed fluxes, i.e., $\int_{\omega_j} v \cdot \mathrm{e}_j \, \mathrm{d}y = H_j$,

where e_j is the unit vector in direction of the cylinder axis. The homogeneous problem $(\mathbb{S}, B) = 0$ has the nontrivial solution $u^\# = (0, 0, 0, 1)$ (see [33, Theorem 5.3]): If $b_j = 0$ in (1.21) then u is a solution to the homogeneous Dirichlet problem with $\nabla v \in L^2(\Omega)$; thus, $v = 0$ and $p=$ const. Now we use the following *generalized Green's formula* for $u = \binom{v}{p}, U = \binom{V}{P} \in \mathbb{D}^l_{\pm\beta} W(\Omega)$ (see [33, Section 5.8.1] and [36, Theorem 4.2]):

$$\begin{aligned}(Su, U)_\Omega - (u, SU)_\Omega + (Nu, V)_{\partial\Omega} - (v, NU)_{\partial\Omega} \\ = \langle \pi_0 u, \pi_1 U\rangle_J - \langle \pi_1 u, \pi_0 U\rangle_J,\end{aligned} \tag{1.24}$$

where $\langle\ ,\ \rangle_J$ is the scalar product in $\mathbb{C}^J$. Let $u \in \mathbb{D}^l_{\pm\beta} W(\Omega)$ be a solution to the Stokes problem $(\mathbb{S}, B)u = (f, g, H)$ with prescribed fluxes. Substituting u and $u^\#$ into (1.24) and using

$$\pi_1 u^\# = 0, \quad \pi_0 u^\# = E \quad \text{with } E = (1, 1, \dots, 1)^\top \in \mathbb{C}^J, \tag{1.25}$$

we obtain the necessary condition

$$\int_\Omega f_4 \,\mathrm{d}x + \int_{\partial\Omega} g \cdot n \,\mathrm{d}o = -\langle \pi_1 u, \pi_0 u^\#\rangle_J = -\sum_{j=1}^J H_j, \tag{1.26}$$

which means that the total flux has to vanish. In [36] it is shown that this condition is also sufficient to obtain a solution $u \in \mathbb{D}^l_{\pm\beta} W(\Omega)$, and this solution is uniquely determined up to a constant in pressure. We summarize this in the following theorem.

THEOREM 1.6 (Stokes system with prescribed fluxes). *Let $l \in \mathbf{N}$, $0 < \beta < \beta^*$, and β^* be defined as above. Then for every $(f, g, H) \in \mathcal{R}^l_\beta W(\Omega, \partial\Omega) \times \mathbb{C}^J$ fulfilling* (1.26) *there exists a solution $u = \binom{v}{p} \in \mathbb{D}^l_{\pm\beta} W(\Omega)$ of the Stokes problem*

$$Su = f \text{ in } \Omega, \quad v = g \text{ on } \partial\Omega, \quad \pi_1 u = H. \tag{1.27}$$

u is uniquely determined by the condition

$$\int_{G_0} p \,\mathrm{d}x = c, \tag{1.28}$$

where $G_0 \subset \Omega$ is an arbitrary nonvoid bounded subdomain and $c \in \mathbb{C}$ is an arbitrary but fixed constant. In this case u obeys the following estimate:

$$\|u; \mathbb{D}^l_{\pm\beta}(\Omega)\| \le C\Big(|H| + \|(f, g); \mathcal{R}^l_\beta W(\Omega, \partial\Omega)\|\Big), \tag{1.29}$$

where the constant C depends on G and on the normalization constant c for the pressure but does not depend on the data (f, g, H).

Remark 1.7. The normalization condition for the pressure is completely voluntary; p may be fixed by prescribing the value of any continuous linear functional Φ on $\mathbb{D}^l_{\pm\beta}(\Omega)$ with $\Phi(u^\#) \neq 0$. For example, for $E \in \mathbb{C}^J$ with $\sum E_j \neq 0$, we can fix p through $\langle E, \pi_0 u\rangle_J = c$.

Theorem 1.6 may also be applied to $(f, g) = (0, 0)$, which leads to a certain characterization of $\ker \mathbb{S}$ with prescribed fluxes. It is clear that linear independent vectors $H^1, \dots, H^k \in \mathbb{C}^J$ lead to linear independent solutions $\eta_1, \dots, \eta_k$ of

$$S\eta = 0 \text{ in } \Omega, \quad \eta = 0 \text{ on } \partial\Omega, \quad \pi_1 \eta = H^j, j = 1, \dots, k, \tag{1.30}$$

provided $\sum_{i=1}^{J} H_i^j = 0$. Let $E = (1, \dots, 1)^\top$ and

$$\mathbb{P} = (\mathcal{P}_1, \dots, \mathcal{P}_J) = \mathbb{I} - J^{-1} E \cdot E^\top, \qquad E \cdot E^\top = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{pmatrix}$$

denote the matrix associated with the projection P on $\mathbb{C}^J$ with $Pc = c - J^{-1}\langle c, E\rangle_J E$. Let η_j denote the solutions of (1.30) where $H^j = \mathcal{P}_j$ is the jth column of the matrix $\mathbb{P}$; this means η_j carries the flux $1 - 1/J$ through the outlet Q_j and the compensating flux is distributed equally through the other $J - 1$ outlets. By Theorem 1.6,

$$\eta_j = \mathcal{U}^1 \cdot \mathcal{P}_j + \mathcal{U}^0 \cdot \mathcal{Q}_j + \tilde{\eta}_j, \tag{1.31}$$

we can fix the vectors $\mathcal{Q}_j$ by $\langle \mathcal{Q}_j, E\rangle_J = 1$.

Then $\eta_1, \dots, \eta_J$ is a basis of $\ker \mathbb{S}$: It is clear that $\operatorname{rank} \mathbb{P} = \dim \operatorname{span}[\mathcal{P}_1, \dots, \mathcal{P}_j] = J-1$. Moreover, due to the special form of $\mathbb{P}$ any $J-1$ columns are linear independent; thus, any $J-1$ elements of $\{\eta_1, \dots, \eta_J\}$ are linear independent and may be completed by $u^\#$ to a basis of $\ker \mathbb{S}$. On the other hand, due to the choice of $\mathcal{Q}_j$, $j = 1, \dots, J$, we have $\sum_{j=1}^{J} \eta_j = u^\#$; thus, $\eta_1, \dots, \eta_j$ also form a basis of $\ker \mathbb{S}$.

We call the matrix $\mathcal{Q}_\Omega = (\mathcal{Q}_1, \dots, \mathcal{Q}_J)$ the *pressure distribution matrix.*

2. Formulation of the approximation problem. We want to approximate the solution of (1.27) by solutions u^R of Stokes problems on bounded subdomains. For $R > R_0$, we cut each outlet at $z_j = R$ and name the resulting domain Ω_R, i.e.,

$$\Omega_R = \{x \in \Omega_0\} \cup \bigcup_{j=1}^{J} \{x \in Q_j : z_j < R\}.$$

For the boundary $\partial\Omega_R$ we have $\partial\Omega_R = (\partial\Omega_R \cap \partial\Omega) \cup \bigcup_{j=1}^{J} \omega_j \times \{R\}$. We call $\partial\Omega(R)$ the intersection of $\partial\Omega_R$ with $\partial\Omega$, $\Gamma_{R,j} = \omega_j \times \{R\}$ the cross section at $z_j = R$, and $\Gamma_R = \bigcup_{j=1}^{J} \Gamma_{R,j}$; thus, $\partial\Omega_R = \partial\Omega(R) \cup \Gamma_R$. Furthermore, we denote the union of edges by $\partial\Gamma_R$. We define the approximation problem in the following way. We fix a system of cut-off functions $\chi_R(x)$ with $\chi_R(x) = 1$ on Ω_0; for each outlet Q_j, $\chi_R = \chi_R(z_j)$ in local coordinates, and

$$\chi_R(z_j) = 1 \text{ for } z_j \le R - 1, \quad \chi_R(z_j) = 0 \text{ for } z_j \ge R - \frac{1}{2}.$$

We look for $u^R = (v^R, p^R)$ as a solution of

$$Su^R = \chi_R f \text{ in } \Omega_R, \qquad v^R = \chi_R g \text{ on } \partial\Omega(R). \tag{2.1}$$

Then we have $Su^R = 0$ in a neighborhood of the edge. As we will see later, this will help us to increase the smoothness properties of the approximating solutions. We choose the boundary condition on the cut cross section $\Gamma_{R,j}$ in dependence of the main asymptotic term $u^{j1}H_j$, where, as before, H_j denotes the flux of the solution u through the outlet Q_j, and u^{j1} coincides with the Poiseuille flow for z_j large enough, i.e.,

$$B^R u^R = \mathcal{H}(H) \quad \text{on } \Gamma_R. \tag{2.2}$$

In the following section we investigate the formal asymptotic behavior of $u - u^R$ for the Dirichlet operator $Du^R = v^R$, for the Neumann boundary operator $Nu^R = \nabla v^R \cdot \nu - p\nu$, and for the operator $(Fu^R)^\top = (\tau_1^\top \cdot Tu^R \cdot \nu, \tau_2^\top \cdot Tu^R \cdot \nu, v^R \cdot \nu)$, where $T = \nabla v + (\nabla v)^\top - p\mathbb{I}$ is the stress tensor ($\mathbb{I}$ the unit matrix in $\mathbf{R}^3$) and τ_1, τ_2 are linear independent tangential vectors on the cross section. This means in local coordinates (y_j, z_j) we choose $\tau_1^\top = (1,0,0)$ and $\tau_2^\top = (0,1,0)$, while $\nu^\top = (0,0,1)$. F is a combination of the Dirichlet operator and a certain Neumann boundary operator, and is usually related to free boundary problems. We call this boundary operator the mixed boundary operator. While calculating the formal asymptotics we neglect the difficulties arising from the edge. We will prove the existence of u^R in suitable function spaces in section 4.

3. Formal asymptotics.

3.1. The general scheme. For the moment we assume that the boundary $\partial\Omega$ is smooth, the data (f, g) are smooth with compact support, and condition (1.26) holds for $H \in \mathbb{C}^J$. We choose R_0 large enough such that $\operatorname{supp} f \subset \Omega_{R_0}$ and $\operatorname{supp} g \subset \partial\Omega(R_0)$. With the transformation $\tilde{u} = u - \mathcal{U}^1 \cdot H$ we pass to the case of zero fluxes; hence, we may assume that $H = 0$ and

$$\int_\Omega f_4 \,\mathrm{d}x + \int_{\partial\Omega} g \,\mathrm{d}o = 0. \tag{3.1}$$

The condition at infinity changes to $\pi_1 u = 0$. By Theorem 1.6 we obtain a solution u of (1.1) and $\pi_1 u = 0$ which is unique under the condition (1.28). By (1.10) we attain in every fixed outlet Q_j for sufficiently large z the asymptotic expansion in exponential solutions of the homogeneous Stokes system

$$u(y,z) \simeq \sum_{\lambda_\mu} \sum_{k,m} c_\mu^{k,m}(f,g) u_{\lambda_\mu}^{(k,m)}(y,z), \tag{3.2}$$

where the sum is taken over all eigenvalues of the $\mathfrak{S}(\lambda)$, where $\operatorname{Im}\lambda \geq 0$, $u_{\lambda_\mu}^{(k,m)}(y,z)$ are the exponential solutions of order k belonging to the eigenvalue λ_k according to (1.8). If we order the eigenvalues according to the size of their imaginary parts β_k, then $\lambda_0 = 0$ and $c_0^{1,1} = 0$ due to the condition $\pi_1 u = 0$; moreover, $c_0^{0,1} = a_j$ according to (1.21). Let $R > R_0 + 1$ and u^R be a solution of the approximation problem. We suppose that u^R admits an asymptotic expansion into solutions of two-limit problems. We fix $x \in Q_j$, $x = (y,z)$ and put $\xi = (y,\zeta) = (y, z-R)$. Then

$$\begin{aligned} u^R(x) &= u^{(0)}(x) + \chi_j(z) U^{(0)}(\xi) + \mathcal{F}^{(1)}(R)\Big(u^{(1)}(x) + \chi_j(z) U^{(1)}(\xi)\Big) + \dots, \\ &\qquad |\mathcal{F}^{(1)}(R)| = O(e^{(-\beta_1 R)} R^{q_{max}}), \end{aligned} \tag{3.3}$$

where $u^{(0)}$ solves $Su^{(0)} = f$ in Ω, $v^{(0)}|_{\partial\Omega} = g$. q_{max} is determined by the maximal length of all Jordan chains belonging to the eigenvalues λ with $\operatorname{Im}\lambda = \beta_1$. $u^{(k)}$, $k = 1, 2, \dots$ are solutions of the *first limit problem*

$$Sw = \mathfrak{f} \quad \text{in } \Omega, \qquad w'|_{\partial\Omega} = 0. \tag{3.4}$$

$U^{(k)}$ are solutions of the *second limit problem*

$$\begin{aligned} &S_\xi W(\xi) = 0 \quad \text{in } \Pi_j^- = \omega_j \times (-\infty, 0), \quad W' = 0 \text{ on } \partial\omega_j \times (0,\infty) \\ &B_\xi W(\xi) = \mathcal{H} \text{ for } \zeta = 0 \end{aligned} \tag{3.5}$$

with suitable data $\mathfrak{f}$, $\mathcal{H}$. In both cases the superscript "$'$" indicates the velocity part of the solutions, i.e., the first three components.

We explain the main idea how to fix the first terms in (3.3) for the case $J = 1$, i.e., Ω has only one outlet. We choose $u^{(0)} \in \mathbb{D}^l_{\pm\beta} W(\Omega)$ as a solution to the Dirichlet problem (1.1) in Ω. By the transformation $\xi = (y, \zeta)$ with $\zeta = z - R$ we pass to R-independent coordinates and calculate the main asymptotic term of $B_\xi u^{(0)} = -\mathcal{H}^{(0)}$ for $\zeta = 0$ by means of (3.2) (where the first coefficients now need not vanish). We remove this term by putting $u^R = u^{(0)} + \chi(z)U^{(0)}(y, z - R) + \cdots$, where $U^{(0)}$ solves (3.5) with $\mathcal{H} = -\mathcal{H}^{(0)}$ and decays exponentially as $\zeta \to -\infty$. $\chi(z)$ is a cut-off function vanishing for $z > R_0$.

Here we use the fact that the assertions before Theorem 1.6 also hold for the problem (3.5) (see section 5 below). This means, in particular, that the solution of (3.5) can always be found in a class of functions growing exponentially to infinity with $\zeta \to -\infty$. If this growth is not too fast, then (up to multiplication with constants) there exists only one nontrivial solution to the homogeneous problem (3.5). Hence, to find $U^{(0)}$ which decays exponentially as $\zeta \to -\infty$, it is sufficient that $\mathcal{H}^{(0)}$ fulfills one compatibility condition arising from the corresponding Green's formula. This can be achieved by the correct choice of $u^{(0)}$.

It is clear that $v^{(0)} + \chi(z)V^{(0)}(y, z - R) = g$ on $\partial\Omega(R)$. We have

$$S(u^{(0)}(y, z) + \chi(z)U^{(0)})(y, z - R) = f + \tilde{f},$$

where $\tilde{f} = [S, \chi]U^{(0)}$ has a compact support. For $U^{(0)}$ we can again use the expansion (3.2), where the coefficients for $\lambda = 0$ now vanish. Using formula (1.8) for $u^{(k,m)}_{\lambda_\mu}$ gives

$$\tilde{f}(y, z) = e^{(i\lambda_1 R)} f^{(1)}(y, z, R) + O\Big(e^{-(\operatorname{Im}\lambda_2 + \varepsilon)}\Big),$$

and $f^{(1)}$ has the structure $f^{(1)}(y, z, R) = \sum_{q=1}^{q_{max}} R^q f^{(1,q)}(y, z)$. We remove these terms by solutions $u^{(1,q)}$ of (3.4) with $\mathfrak{f} = -f^{(1,q)}$, then calculate again the leading term of $B(u^{(0)} + \chi U^{(0)} + e^{i\lambda_1 R} \sum_q R^q u^{(1,q)})$ for $z = R$, and so on.

If $J > 1$, one has to take into account that $u^{(k)}$ influences all outlets, and the eigenvalues λ_μ for $\lambda \neq 0$ may be different in all outlets. Important for the behavior of $u - u^R$ is the choice of $u^{(0)}$ and the first step in the calculations, which can be done simultaneously in all outlets, since the first eigenvalue, $\lambda = 0$, is the same for all outlets.

3.2. The Dirichlet condition and the mixed boundary condition on the artificial boundary. Let f, g, and u be as in section 3.1. For sufficiently large R let $u^R = (v^R, p^R)$ be a solution to $Su^R = f$ in Ω_R, $v^R = g$ on $\partial\Omega(R)$, and $v^R = 0$ on Γ_R. We outline that u^R exists due to (3.1). We choose $u^{(0)} = u$. To calculate $v^{(0)}|_{z_j = R}$, we use (3.2) and apply the transformation $z = \zeta + R$ to the exponential solutions $u^{(k,m)}_\lambda(y, z)$. With (1.8) it follows that

$$u^{(k,m)}_\lambda(y, z) = e^{i\lambda R} \sum_{q=0}^{k} \frac{1}{q!} (iR)^q u^{(k-q,m)}_\lambda(y, \zeta).$$

For fixed j we obtain

$$v(y, R) = \sum_{\operatorname{Im}\lambda = \beta_1} e^{i\lambda R} \mathcal{M}^{(k,m)}_\lambda(f, g, R) v^{(k,m)}_\lambda(y, 0) + O(e^{(-\beta_2 + \varepsilon)R}).$$

Here the sum is taken over all eigenvalues of $\mathfrak{S}(\lambda)$ on the line $\operatorname{Im}\lambda = \beta_1$, $\beta_1 > 0$. Each $\mathcal{M}_\lambda^{(k,m)}(f,g,R)$ is a polynomial in R containing also the coefficients $c_\mu^{k,m}(f,g)$ of the representation (3.3). Each term in the sum has the form

$$Ce^{-\beta_1 R}\, e^{i\operatorname{Re}\lambda R} R^l v_\lambda^{k_l,m}(y,0) =: e^{-\beta_1 R} R^l \mathcal{H}^{(1,l)}(y) \tag{3.6}$$

with $\beta_1 = \operatorname{Im}\lambda > 0$. Applying the scheme of 3.1, we see that $U^{(0)} = 0$ in each outlet. Moreover, we find solutions $U^{(1,l)}$ to the second limit problem with $\mathcal{H} = \mathcal{H}^{(1,l)}$ decaying exponentially to 0 as $\zeta \to -\infty$: We recall that there exists only one nontrivial solution $U = \binom{V}{P}$ to the homogeneous problem (3.5) with

$$\sup_{\zeta\to-\infty} e^{\beta\zeta} U(y,\zeta) < \infty, \quad 0 < \beta < \beta^*.$$

Then it is clear that this solution $U = (0,0,0,1)^\top = u^\#$. Green's formula (1.7) in Π_j^- leads to the necessary condition

$$\int_{\omega_j} \mathrm{e}_z^\top \cdot H^{(1,l)}(y)\,\mathrm{d}y = 0, \tag{3.7}$$

where $\mathrm{e}_z = (0,0,1)^\top$ is the unit vector in z-direction. Replacing Ω in the same formula by $\omega_j \times [0,t]$, $u = u_\lambda^{(k,m)}$, $U = u^\#$ and passing with $t \to \infty$, we see $\int_{\omega_j} \mathrm{e}_z^\top v_\lambda^{(k,m)}(y,0)\,\mathrm{d}y = 0$ if $\operatorname{Im}\lambda > 0$; thus (3.7) holds.

Now we proceed as in section 3.1 and obtain the formal asymptotic error estimate

$$|u(x) - u^R(x)| = O(e^{-(\beta^*-\varepsilon)R}), \quad \varepsilon > 0 \text{ arbitrary},$$

if we keep in mind that β^* is the mimimum of all imaginary parts of the nonzero eigenvalues in all outlets.

For $Bu^R = Fu^R$ we have to observe the following Green's formula ($Tu = \nabla v + (\nabla v)^\top - p\mathbb{I}$):

$$\begin{aligned}&\int_{\Pi_j^-} S_1 u \cdot U\,\mathrm{d}x + \int_{-\infty}^0 \int_{\partial\omega_j} Tu\cdot\nu\cdot V\,\mathrm{d}o + \int_{\omega_j} Fu(y,0)\cdot F_0 U(y,0)\,\mathrm{d}y\\ &= \int_{\Pi_j^-} u\cdot S_1 U\,\mathrm{d}x + \int_{-\infty}^0 \int_{\partial\omega_j} TU\cdot\nu\cdot v\,\mathrm{d}o + \int_{\omega_j} F_0 u(y,0)\cdot FU(y,0)\,\mathrm{d}y\end{aligned} \tag{3.8}$$

with

$$S_1 u = \begin{pmatrix} -\Delta v - \nabla\operatorname{div} v + \nabla p \\ -\operatorname{div} v \end{pmatrix}, \quad F_0 u = (v_\tau, -\nu^\top \cdot Tu\cdot\nu).$$

The solution to the homogeneous second limit problem is again $u^\#$; thus Green's formula (3.8) leads to the same compatibility conditions, and we may use analogous calculations as for the Dirichlet operator.

3.3. The Neumann condition on the artificial boundary. Now let $u^R = \binom{v^R}{p^R}$ be a solution of

$$Su^R = f \text{ in } \Omega_R, \quad v^R = g \text{ on } \partial\Omega(R), \quad Nu^R = (\nabla v^R - p\mathbb{I})\cdot\nu = 0 \text{ on } \Gamma_R.$$

Let $j \in \{1, \dots, J\}$ and $x = (y, z) \in Q_j$ be fixed with $z > R_0$. If we put $u^{(0)} = u$, then to apply the scheme developed above, we use the expansion (3.2) and achieve

$$u(x) = \mathcal{U}^0 \cdot a + \sum_{\operatorname{Im}\lambda=\beta_1} e^{-\beta_1 z} \cdots + \cdots \tag{3.9}$$

with $a = (a_1, \dots, a_J) = \pi_0 u$. For $z = R$ this leads to

$$Nu(y, R) = \partial_z v(y, R) - p(y, R)\mathrm{e}_z = a_j + O(e^{-(\beta_1 - \varepsilon)R})$$

with some arbitrary small but in general positive ε. In order to remove the term a_j on $\Gamma_{R,j}$, $U^{(0)} = (V^{(1)}, P^{(1)})$ has to be a solution of (3.5) with

$$N_\xi U^{(0)}(\xi) = a_j \tag{3.10}$$

with $\xi = (y, z - R)$; (y, z) are the local coordinates in the outlet Q_j. If we choose $U^{(0)}(\xi) = u^{j0}(\xi)a_j$ in every outlet, then extending $\chi_j U^{(0)}$ by zero to $x \notin Q_j$ and substituting $\chi_j U^{(0)}$ into formula (3.3), we see that $u^{(1)}$ has to solve (3.4) with $\mathfrak{f} = \sum_j S(\chi_j U^{(0)})$. This gives $u^{(1)} = \mathcal{U}^0 \cdot a$ and we arrive again at the second limit problem with condition (3.10).

For this reason we have to find $U^{(0)}(\xi)$ in a class of functions where the velocity part $V^{(0)}$ and the pressure part $P^{(1)}$, together with all derivatives, exponentially die to 0 as $\zeta \to -\infty$, i.e.,

$$\sup_{\zeta<-1} e^{-\beta\zeta} |\partial^\alpha U^{(0)}(y, \zeta)| < \infty \qquad \text{for } 0 < \beta < \beta^*,$$

where β^* has the same meaning as in 1.3. Using Green's formula (1.7) in Π_j^-, we find that this implies

$$\int_{\omega_j(0)} NU^{(0)} \cdot V \, \mathrm{d}o = \int_{\omega_j} a_j V_3(y, 0) \, \mathrm{d}y = 0 \tag{3.11}$$

for any solution $U = (V, P)$ of

$$SU = 0 \text{ in } \Pi_j^-, \qquad V = 0 \text{ on } \partial\omega_j \times (-\infty, 0), \qquad NU = 0 \text{ for } \zeta = 0 \tag{3.12}$$

with $V \in H^2_{loc}(\overline{\Omega})$, $P \in H^1_{loc}(\overline{\Omega})$, and

$$\sup_{\zeta<-1} e^{\beta'\zeta} |\partial^\alpha V(y, \zeta)| < \infty \tag{3.13}$$

for some $\beta' < \beta$, $\alpha \in \mathbf{N}_0^3$. As already mentioned, up to multiplication with constants there exists only one nontrivial solution to (3.12) and (3.13), which is the Poiseuille flow in this case. From this result we see that only for $J = 1$, i.e., if the domain has only one outlet, can (3.11) be fulfilled through the change of the normalization condition (1.28) to $a_1 = 0$; whereas for $J \geq 2$, only $\sum_{j=1}^J a_j = 0$ is possible according to Remark 1.7. Thus we change $u^{(0)}$ to $u^{(0)} = u + \eta \cdot c$; $c = c(R) \in \mathbb{C}^J$ has to be chosen properly; $\eta(x)$ is the $4 \times J-$matrix with columns η_j, where η_j are the special solutions of the homogeneous Stokes problem defined in Remark 1.7. Using the representation (3.9) for u and (1.31) for η, we obtain

$$u^{(0)} = \mathcal{U}^1 \cdot \mathbb{P} \cdot c + \mathcal{U}^0 \cdot (a + \mathcal{Q}_\Omega \cdot c) + \tilde{u}, \tag{3.14}$$

where $\tilde{u}$ decays exponentially. To calculate the leading term of $Nu^{(0)}$ on Γ_R, we fix one outlet Q_j and rewrite (3.14) in ξ-coordinates, $\xi = (y, z - R)$; this gives

$$u^{(0)}(\xi) = \left[u^{j1}(\xi)(\mathbb{P}\cdot c)_j + u^{j0}\left(a_j + (\mathcal{Q}_\Omega \cdot c)_j - 2\varpi_j R(\mathbb{P}\cdot c)_j\right)\right] + \tilde{\tilde{u}}\,.$$

The term $-2\varpi_j R(\mathbb{P}\cdot c)_j$ appears because $u^{j1}(y,z) = u^{j1}(\xi) - 2\varpi_j(0,0,0,R)$. So the main discrepancy at $\xi = 0$ is produced by $N[\ldots]$, and we look for $U^{(0)}$ as a solution of (3.5) whereas the boundary condition (3.10) changes to

$$N_\xi U^{(0)}(\xi) = -N_\xi\left[u^{j1}(y,0)(\mathbb{P}\cdot c)_j + u^{j0}(a_j + (\mathcal{Q}_\Omega\cdot c)_j - 2\varpi_j R(\mathbb{P}\cdot c)_j)\right] =: \mathcal{H}^j_{(1)}(y).$$

The boundary operator annuls the first term on the right-hand side, hence $N_\xi u^{(0)}(\xi) = 0$ if $a_j + (\mathcal{Q}_\Omega\cdot c)_j - 2\varpi_j R(\mathbb{P}\cdot c)_j = 0$. This condition has to be accomplished in every outlet Q_j, so we arrive at a linear system for $c \in \mathbb{C}^J$:

$$-\mathcal{Q}_\Omega\cdot c + 2\varpi_j R\mathbb{P}\cdot c = a. \tag{3.15}$$

To show the solvability of this system, we substitute

$$c = \mathbb{P}\cdot c + c_E E \text{ with } c_E = J^{-1}\langle c, E\rangle_J, \quad E = (1,\ldots,1)^\top \tag{3.16}$$

into (3.15), remembering the choice of $\mathcal{Q}_j$; this leads to

$$-\mathcal{Q}_\Omega\cdot\mathbb{P}\cdot c - c_E E + 2\varpi_j R\mathbb{P}\cdot c = a. \tag{3.17}$$

We apply $\mathbb{P}$ to this system and obtain

$$-\mathbb{P}\cdot\mathcal{Q}_\Omega\cdot\mathbb{P}\cdot c + 2\varpi_j R\mathbb{P}\cdot c = \mathbb{P}a$$

because $\mathbb{P}^2 = \mathbb{P}$. Since $\mathbb{P}$ gives the identity on $\mathbb{P}\mathbb{C}^J$, this system is uniquely solvable in $\mathbb{P}\mathbb{C}^J$ for sufficiently large R. Now we multiply (3.17) scalar by E. Using the symmetry of $\mathcal{Q}_\Omega$ together with (3.16) we calculate $c_E = -J^{-1}\langle a, E\rangle_J$, and thus prove the unique solvability of (3.15).

We have

$$c_E = const, \quad \mathbb{P}c(R) = O(R^{-1}) \qquad \text{as } R\to\infty. \tag{3.18}$$

With this choice of c we find $U^{(0)}$ in each outlet such that

$$S\Big(\chi_j(z)U^{(0)}\Big)(y, z-R) = O(e^{-\beta_1 R}R^{q(j)})$$

and we proceed as indicated in section 3.1. As a result, we obtain

$$|u(x) - u^R(x)| = |\eta\cdot c(R)(x)| + O(e^{-(\beta^*-\varepsilon)R}) = O(1).$$

If the normalization condition $\langle a, E\rangle_J = 0$ is chosen, then the previous calculations lead to $c_E = 0$ and $|u - u^R(x)| = O(R^{-1})$.

4. Solution of the approximation problem. The aim of the following sections is to prove the existence of solutions u^R to the approximation problem and the error estimate for $u - u^R$ which justifies the formal asymptotic estimates rigorously. We carry out the proofs for the mixed boundary condition on the artificial boundary Γ_R, since this condition gives the best regularity properties of the approximating solutions (see Theorem 4.4 below).

For the approximation problem itself, it is possible to prove the existence of a solution $u^R \in \mathcal{D}^l H(\Omega_R) = H^{l+1}(\Omega_R)^3 \times H^l(\Omega_R)$ with comparably elementary methods. However, to obtain an error estimate for the difference $u - u^R$, it is necessary to prove the existence of solutions and uniform estimates for the problem

$$SU = f \text{ in } \Omega_R, \tag{4.1}$$

$$V = g \text{ on } \partial\Omega(R), \qquad FU = h \text{ on } \Gamma_R, \tag{4.2}$$

where (f, g, h) are general data in suitable function spaces. If we look for a solution $U \in H^2(\Omega_R)^3 \times H^1(\Omega_R)$ at least, then f_4, g, h have to fulfill the necessary condition arising from the Gauss theorem:

$$\int_{\Omega_R} f_4 \,\mathrm{d}x + \int_{\partial\Omega(R)} \nu^\top \cdot g \,\mathrm{d}o + \int_{\Gamma_R} \nu^\top \cdot h \,\mathrm{d}o = 0. \tag{4.3}$$

It is clear that if $u \in \mathbb{D}^l_{\pm\beta}(\Omega)$ is a solution to the Stokes problem with prescribed fluxes and $u^R \in \mathcal{D}^l H(\Omega_R)$ is a solution to the approximation problem, then the difference $U = u - u^R \in \mathcal{D}^l H(\Omega_R)$ fulfills (4.1), (4.2) with $f' \in H^{l-1}(\Omega_R)^3$, $f_4 \in H^l(\Omega_R)$, $g \in H^{l+1/2}(\partial\Omega(R))^3$, and $h \in H^{l-1/2}(\Gamma_R)^2 \times H^{l+1/2}(\Gamma_R)$. (The notation for h must be read in local coordinates of the outlet Q_j.) On the other hand, it is obvious that even for $l = 1$ we cannot obtain a solution $U \in \mathcal{D}^1 H(\Omega_R)$ of (4.1), (4.2) for arbitrary data $(f, g, h) \in \mathcal{R}^1 H(\Omega_R, \partial\Omega(R), \Gamma_R)$, even if (4.3) is satisfied. In a 3-dimensional domain the velocity part V of this solution would be continuous up to the boundary, and hence also on the edges of Ω_R, which would force additional compatibility conditions for the data. Thus, the general problem (4.1), (4.2) has to be treated in weighted spaces with weights $\sim \rho^\gamma$, where $\rho(x) = \mathrm{dist}(x, \partial\Gamma_R)$ in the vicinity of $\partial\Gamma_R$.

Since $\partial\Gamma_R$ is a disjoint union of 1-dimensional C^{l+2}-manifolds, we find $\varepsilon_0 > 0$ such that $\mathrm{dist}(x, \partial\Gamma_R)$ is of class C^{l+2} on $\mathcal{O}_{\varepsilon_0}(\Gamma_R) = \{x, \mathrm{dist}(x, \Gamma_R) < \varepsilon_0\}$. We choose a basic weight function $\rho \in C^{l+2}(\Omega_R)$ with $\rho(x) = \mathrm{dist}(x, \partial\Gamma_R)$ for $x \in \mathcal{O}_{\varepsilon_0} \cap \Omega_R$, and $\rho(x) = 1$ for $x \in \Omega_{R-2\varepsilon_0}$ and introduce spaces of Kondratiev's type on Ω_R.

Definition 4.1. *Let $l \in \mathbf{N}_0$, $\gamma \in \mathbf{R}$, $\varphi \in C_0^\infty(\overline{\Omega}_R \backslash \partial\Gamma_R)$. We set*

$$\|\varphi; V^l_\gamma(\Omega_R, \partial\Gamma_R)\| = \sum_{|\alpha| \le l} \left(\|\rho^{\gamma - l + |\alpha|} \partial^\alpha \varphi; L^2(\Omega_R)\|^2 \right)^{1/2} \tag{4.4}$$

and $V^l_\gamma(\Omega_R, \partial\Gamma_R)$ the closure of $C_0^\infty(\overline{\Omega}_R \backslash \partial\Gamma_R)$ in the norm (4.4).

By $V^{l-1/2}_\gamma(\partial\Omega(R), \partial\Gamma_R)$ and $V^{l-1/2}_\gamma(\Gamma_R, \partial\Gamma_R)$ we denote the spaces of traces on $\partial\Omega(R)$ and Γ_R. We define

$$\|\varphi; V^{l-1/2}_\gamma(M, \partial\Gamma_R)\| = \inf \|\tilde{\varphi}; V^l_\gamma(\Omega_R, \partial\Gamma_R)\|$$

for $M = \partial\Omega(R)$ and $M = \Gamma_R$, where the infimum is taken over all $\tilde{\varphi} \in V^l_\gamma(\Omega_R, \partial\Gamma_R)$ with $\tilde{\varphi}|_M = \varphi$.

$V^l_\gamma(\Omega_R, \partial\Gamma_R)$ coincides with the space of all functions

$$\{\varphi \in H^l_{loc}(\Omega_R); \|\varphi; V^l_\gamma(\Omega_R, \partial\Gamma_R)\| < \infty\}$$

and thus is a space of regular distributions on Ω_R. The trace space $V^{l-1/2}_\gamma(\Gamma_R, \partial\Gamma_R)$ is the union of the traces on the J cross sections $\Gamma_{R,j} = \Gamma_R \cap Q_j$, where the weight

$\rho(x) = \operatorname{dist}(x, \partial\omega_j \times \{R\})$ in the ε_0-neighborhood of $\partial\omega_j \times \{R\}$. Thus we denote the trace spaces on the cross section $\omega_j \times \{R\}$ by $V_\gamma^{l-1/2}(\omega_j(R), \partial\omega_j)$.

Analogous to the previous section we define the natural domain and range of the problem (4.1), (4.2),

$$\begin{aligned}
\mathcal{D}_\gamma^l V(\Omega_R) &= V_\gamma^{l+1}(\Omega_R, \partial\Gamma_R)^3 \times V_\gamma^l(\Omega_R, \partial\Gamma_R),\\
\mathcal{R}_\gamma^l V(\Omega_R) &= V_\gamma^{l-1}(\Omega_R, \partial\Gamma_R)^3 \times V_\gamma^l(\Omega_R, \partial\Gamma_R),\\
\mathcal{R}_\gamma^l V(\Omega_R, \partial\Omega(R), \Gamma_R) &= \mathcal{R}_\gamma^l V(\Omega_R) \times V_\gamma^{l+1/2}(\partial\Omega(R), \partial\Gamma_R)^3\\
&\quad \times \prod_{j=1}^J V_\gamma^{l-1/2}(\omega_j(R), \partial\omega_j)^2 \times V_\gamma^{l+1/2}(\omega_j(R), \partial\omega_j).
\end{aligned} \tag{4.5}$$

Then the operator

$$\begin{aligned}
(S, D, F) : \mathcal{D}_\gamma^l V(\Omega_R) &\to \mathcal{R}_\gamma^l V(\Omega_R, \partial\Omega(R), \Gamma_R),\\
u &\mapsto (Su, v|_{\partial\Omega(R)}, Fu|_{\Gamma_{R,j}}, j = 1, \dots, J)
\end{aligned} \tag{4.6}$$

defines a continuous linear operator. Here the notation

$$Fu|_{\Gamma_{R,j}} \in V_\gamma^{l-1/2}(\omega_j(R), \partial\omega_j(R))^2 \times V_\gamma^{l+1/2}(\omega_j(R), \partial\omega_j(R))$$

has to be understood in the local coordinates of Q_j:

$$\partial_{y_i} v_3 - \partial_z v_i \in V_\gamma^{l-1/2}(\omega_j(R), \partial\omega_j(R)) \text{ and } v_3 \in V_\gamma^{l+1/2}(\omega_j(R), \partial\omega_j(R)).$$

To derive Fredholm properties of the mapping (4.6) at least for γ in a certain bounded interval, we have to combine results from the general theory of elliptic systems in domains with smooth edges with results which are already known for the problem (4.1), (4.2).

THEOREM 4.2. *Let $l \in \mathbf{N}$, $\partial\Omega \in C^{l+2}$, $\gamma \in \mathbf{R}$. The operator (S, D, F) defined by (4.6) is Fredholm for $|\gamma - l| < 1$. For $(f, g, h) \in \mathcal{R}_\gamma^l V(\Omega_R, \partial\Omega(R), \Gamma_R)$ there exists a solution to the problem (4.1), (4.2) iff*

$$\int_{\Omega_R} f_4 \,\mathrm{d}x + \int_{\partial\Omega(R)} \nu^\top \cdot g \,\mathrm{d}o + \int_{\Gamma_R} \nu^\top \cdot h \,\mathrm{d}o = 0. \tag{4.7}$$

The solution is unique up to a constant in pressure and thus obeys the following estimate:

$$\|u; \mathcal{D}_\gamma^l V(\Omega_R)\| \le C\Big(\|(f, g, h); \mathcal{R}_\gamma^l V(\Omega_R, \partial\Omega(R), \Gamma_R)\| + \|p; L^2(G_0)\|\Big), \tag{4.8}$$

where G_0 is an arbitrary nonvoid subdomain of Ω_R and $C = C(R, \Omega, l)$ is a constant.

Proof. The domain Ω_R is a bounded domain with a smooth edge $\partial\Gamma_R$. This means for every $x_0 \in \partial\Gamma_R$ there exists a neighborhood $\mathcal{O}(x_0)$ and a C^{l+2}- diffeomorphism $\Phi : \mathcal{O} \to \mathcal{O}'(0)$ ($\mathcal{O}'(0)$ is a neighborhood of 0) such that $\Phi(\mathcal{O} \cap \Omega_R) = \mathcal{O}'(0) \cap \mathbb{D}$, $\mathbb{D} = d_{\pi/2} \times \mathbf{R}$ is an infinite wedge, $d_{\pi/2} = \{x = (r, \theta) \in \mathbf{R}^2 : 0 < r < \infty, 0 < \theta < \pi/2\}$. The treatise of the boundary value problem (4.1), (4.2) in Ω_R is connected to the analysis of the model problem in the wedge $\mathbb{D}$ which takes the following form:

$$Su(x) = f(x) \quad \text{for } x \in \mathbb{D} = d_{\pi/2} \times \mathbf{R}, \tag{4.9}$$

$$Du(x) = v(x) = g(x_2, x_3) \quad \text{for } x_1 = 0, \tag{4.10}$$

$$Fu = \left(\frac{\partial v_2}{\partial x_1} + \frac{\partial v_1}{\partial x_2}, -v_2, \frac{\partial v_3}{\partial x_2}\right)(x) = h(x_1, x_3) \quad \text{for } x_2 = 0. \tag{4.11}$$

We set $\Gamma^D = \{0\} \times \mathbf{R}^+ \times \mathbf{R}$, which corresponds to the lateral surface $\partial\Omega(R)$, and $\Gamma^F = \mathbf{R}^+ \times \{0\} \times \mathbf{R}$, which corresponds to the artificial boundary Γ_R. The union of edges, $\partial\Gamma_R$, corresponds to the infinite edge $M' = \{0\} \times \{0\} \times \mathbf{R}$.

In $\mathbb{D}$ let $\rho(x) = (x_1^2 + x_2^2)^{1/2}$, i.e., for all $x \in \mathbb{D}$, $\rho(x)$ is the distance of x to the edge M'; let $V_\gamma^l(\mathbb{D}, M')$ be defined as in Definition 4.1. We introduce the natural domain and range $\mathcal{D}_\gamma^l V(\mathbb{D})$ and $\mathcal{R}_\gamma^l V(\mathbb{D}, \Gamma^D, \Gamma^F)$ in an analogous way as in (4.5). Then the operator $(S, D, F) : \mathcal{D}_\gamma^l V(\mathbb{D}, M') \to \mathcal{R}_\gamma^l V(\mathbb{D}, \Gamma^D, \Gamma^F)$ defines an isomorphism for $|\gamma - l| < 1$. This was proved for $0 \le \gamma - l < 1$ in [49] and [32]. For $-1 < \gamma - l < 0$ it follows from Proposition 8.2.8 of [33] together with [49, pp. 353, 402] concerning the eigenvalues of the associated 2-dimensional problems in the angle $d_{\pi/2}$.

Theorem 8.3.1 in [33] now gives the Fredholm property of the mapping (4.6) for $|\gamma - l| < 1$. Moreover, from the proof of this theorem it follows that kernel and cokernel of (4.6) are independent of γ and l in this interval. Since $Su = f$ iff

$$S_1 u := (-\Delta v - \nabla \operatorname{div} v + \nabla p, -\operatorname{div} v) = (f' + \nabla f_4, f_4),$$

the same is valid for the operator

$$\mathbb{S}_\gamma^l : \mathcal{D}_\gamma^l V(\Omega_R) \ni u \to (S_1 u, v|_{\partial\Omega(R)}, Fu|_{\Gamma_R}) \in \mathcal{R}_\gamma^l V(\Omega_R, \partial\Omega(R), \Gamma_R).$$

The operator (S_1, D, F) is formally self-adjoint with respect to Green's formula

$$\begin{aligned} &(S_1 u, U)_{\Omega_R} + (v, N_1 U)_{\partial\Omega(R)} + (Fu, F_0 U)_{\Gamma_R} \\ &= (u, S_1 U)_{\Omega_R} + (N_1 u, V)_{\partial\Omega(R)} + (F_0 u, FU)_{\Gamma_R}, \end{aligned} \tag{4.12}$$

where $N_1 u = Tu \cdot \nu$ and, as in section 3.2, $(F_0 u)^\top = (v_\tau, -\nu^\top \cdot N_1 u)$. (4.12) is valid for $u \in \mathcal{D}_\gamma^l V(\Omega_R)$ and $U \in \mathcal{D}_{2l-\gamma}^l V(\Omega_R)$. (We note that for $|\gamma - l| < 1$ also $|(2l - \gamma) - l| < 1$ holds.) Therefore, Theorem 8.3.3 in [33] leads to $\operatorname{Ind} \mathbb{S}_\gamma^l = 0$, and it remains to calculate the kernel of $\mathbb{S}_\gamma^l$ for one exponent γ such that $|\gamma - l| < 1$, for example, for $\gamma = l$. We fix $u = (v, p) \in \mathcal{D}_l^l V(\Omega_R)$ such that $(S_1, D, F)u = (S, D, F)u = 0$. In this case we have $\nabla v \in L^2(\Omega_R)$ and $p \in L^2(\Omega_R)$. Since $v = 0$ on $\partial\Omega(R)$ we may apply Poincaré's inequality and obtain $v \in H^1(\Omega_R)$. Multiplying the equation $S_1 u = 0$ scalar (in $L^2(\Omega_R)$) by u, then using integration by parts and the boundary conditions, leads to $v = 0$ and $p = \text{const}$. Since the function $p \equiv 1$ is contained in V_γ^l for all γ with $\gamma - l > -1$, we have $\ker \mathbb{S}_\gamma^l = \{(0, 0, 0, c) : c \in \mathbf{R}\}$. Now the necessity of condition (4.7) follows from (4.12) and the sufficiency from $\operatorname{Ind} \mathbb{S}_\gamma^l = 0$. □

Remark 4.3. The other boundary conditions on the cut cross sections Γ_R mentioned in section 3 can be treated exactly along the same scheme; only the admissible intervals for the weight index γ are different.

We apply Theorem 4.2 to prove the existence of solutions to the approximation problem.

THEOREM 4.4. *Let $l \in \mathbf{N}$, $\partial\Omega \in C^{l+2}$, let $f \in H^{l-1}(\Omega)^3 \times H^l(\Omega)$, $g \in H^{l+1/2}(\partial\Omega)$, and $H \in \mathbb{C}^J$ be given such that $\int_\Omega |f_4| \, dx + \int_{\partial\Omega} |\nu \cdot g| \, do < \infty$, and the compatibility condition*

$$\int_\Omega f_4 \, dx + \int_{\partial\Omega} \nu^\top \cdot g \, do + \sum_{j=1}^J H_j = 0 \tag{4.13}$$

is fulfilled. Let $\{\chi_R\}_{R>1}$ be a system of cut-off functions with the following properties: $\chi_R(x) = \chi_R(z_j)$ in each outlet Q_j, $\chi_R(x) = 1$ for $x \in \Omega_0$ and for $x = (y_j, z_j)$

with $z_j < R-1$, *and* $\chi_R(y_j,z_j)=0$ *for* $z_j > R-1/2$. *Let* $\mathcal{U}^1$ *be defined as in section* 1.2, *i.e.,* $\mathcal{U}^1 = \binom{\mathcal{V}^1}{\mathcal{P}^1}$, $\mathcal{V}^1$ *is the* $3\times J$*-matrix with rows* $\chi_j v^{j1}$, $j=1,\dots,J$; $\mathcal{P}^1=(\chi_1 p^{j1},\dots,\chi_J p^{J1})$, *and* $u^{j1}=\binom{v^{j1}}{p^{j1}}$ *indicates the Poiseuille flow in the outlet* Qj.

Then there exists a unique solution $u^R=(v^R,p^R)$ *with* $v^R\in H^{l+1}(\Omega_R)^3$, $p^R\in H^l(\Omega_R)$ *to the approximation problem*

$$Su^R = \chi_R f \quad \text{in } \Omega_R \tag{4.14}$$

$$v^R = \chi_R g \quad \text{on } \partial\Omega(R) \tag{4.15}$$

$$Fu^R = F(\mathcal{U}^1\cdot(H+H_\infty)) \quad \text{on } \Gamma_R \tag{4.16}$$

$$\int_{G_0} p^R \,\mathrm{d}o = 0, \tag{4.17}$$

where $H_\infty^\top=(H_{\infty,1},\dots,H_{\infty,J})\in\mathbb{C}^J$ *with*

$$H_{\infty,j}=\int_{Q_j}(1-\chi_R)(z_j)f_4(y_j,z_j)\,\mathrm{d}y_j\,\mathrm{d}z_j+\int_{\partial Q_j}\nu^\top\cdot(1-\chi_R)(z_j)g(y_j,z_j)\,\mathrm{d}o,$$

and $G_0\subset\Omega_0$ *is a nonvoid subdomain such that* $G_0\cap Q_j=\emptyset$ *for* $j=1,\dots,J$.

Remark 4.5. The integrability conditions for f_4 and g look artificial, but they are fulfilled if the data meet the conditions of Theorem 1.6 or, trivially, if $f_4=0$ and $g=0$. Thus we can use this result to approximate the solution $u\in\mathbb{D}^l_{\pm\beta}W(\Omega)$ for exponentially dying data as well as for weak solutions with zero boundary values and $f=\binom{f'}{0}$.

Proof of Theorem 4.4. First of all, it is easy to see that $\chi_R f\in\mathcal{R}^l_\gamma V(\Omega_R)$, $\chi_R g\in V^{l+1/2}_\gamma(\partial\Omega(R))$ for all $\gamma\in\mathbf{R}$, since the support is separated from the edges $\partial\Gamma_R$. For the boundary values on Γ_R we have to use more subtle arguments, which are collected in the following lemma.

LEMMA 4.6. *Suppose* $w\in H^l(\Omega_R)$ *and* $w=0$ *on* $\partial\Omega(R)\cap\mathcal{O}$, *where* $\mathcal{O}$ *is a neighborhood of* $\partial\Gamma_R$. *Then* $w\in V^l_\gamma(\Omega_R)$ *for all* $\gamma-l\geq -1$.

Proof (see also [33, p. 31] for the 2-dimensional case). Because $V^l_{\bar\gamma}(\Omega_R)\supset V^l_\gamma(\Omega_R)$ for $\bar\gamma>\gamma$, it is sufficient to prove $w\in V^l_{l-1}(\Omega_R)$. Since $w\in H^l(\Omega_R)$ we have

$$\|\partial^\alpha w;V^0_{-1+|\alpha|}(\Omega_R)\|\leq C\sup_{x\in\mathcal{O}\cap\Omega_R}\rho(x)^{-1+|\alpha|}\|\partial^\alpha w;L^2(\Omega_R)\|<\infty$$

for all $1\leq|\alpha|\leq l$. It remains to show $w\in V^0_{-1}(\Omega_R)$. We choose a covering $\{\mathcal{O}_\mu\}_{\mu=1}^N$ of Ω_R such that $\bigcup_{\mu=1}^N\mathcal{O}_\mu$ covers the edges $\partial\Gamma_R$ and C^l-diffeomorphisms Φ_μ, such that $\Phi_\mu(\mathcal{O}_\mu)=\mathbb{D}\cap Z(T)$ for some $T>0$. Here $Z(T)=\{(x_2,x_2,x_3);x_1^2+x_2^2<T^2,-T<x_3<T\}$ is a circular cylinder. Moreover, we suppose $\Phi_\mu(\mathcal{O}_\mu\cap\partial\Omega(R))\subset\Gamma^D$ and $\Phi_\mu(\mathcal{O}_\mu\cap\Gamma_R)\subset\Gamma^F$ (see the notations in the proof of Theorem 4.2). Let χ_μ be a partition of unity subordinated to the covering $\{\mathcal{O}_\mu\}_\mu$. Then $w\in V^0_{-1}(\Omega_R)$ is equivalent to $\tilde w:=\chi_\mu w\circ\Phi_\mu^{-1}\in V^0_{-1}(\mathbb{D})$ for every μ. We have $\chi_\mu w\circ\Phi_\mu^{-1}\in H^1(\mathbb{D}\cap B(0,T))$, moreover, $\chi_\mu w\circ\Phi_\mu^{-1}=0$ on Γ^D. We use polar coordinates (ρ,θ) on $d_{\pi/2}$, then with Friedrichs' inequality we arrive at

$$\begin{aligned}\int_{\mathbb{D}\cap Z(T)}|\nabla\tilde w|^2\,\mathrm{d}x&\geq\int_0^T\int_0^T\int_0^{\pi/2}|\partial_\theta\tilde w|^2\rho^{-1}\,\mathrm{d}\theta\,\mathrm{d}\rho\,\mathrm{d}x_3\\&\geq\int_0^T\int_0^T\int_0^{\pi/2}|\tilde w|^2\rho^{-1}\,\mathrm{d}\theta\,\mathrm{d}\rho\,\mathrm{d}x_3=\int_{\mathbb{D}\cap Z(T)}|\tilde w|^2\rho^{-2}\,\mathrm{d}x,\end{aligned}$$

which proves the lemma. □

Using this result, we obtain $\mathcal{U}^1 \cdot A|_{\Omega_R} \in \mathcal{D}_l^l(\Omega_R)$ for every $A \in \mathbb{C}^J$; thus $F(\mathcal{U}^1 \cdot (H + H_\infty))$ is contained in the corresponding trace space. With Theorem 4.2 we obtain a solution $u \in \mathcal{D}_l^l V(\Omega_R)$ to the approximation problem; it is clear then that it is possible to fix the constant in pressure by condition (4.17).

It remains to verify the regularity property. From the definition of $\mathcal{D}_l^l V(\Omega_R)$ we know already that $u^R \in H^{l+1}(\Omega_{R-\varepsilon})^3 \times H^l(\Omega_{R-\varepsilon})$; moreover, $\nabla v^R \in L^2(\Omega_R)$ and $p^R \in L^2(\Omega_R)$. It remains to treat a neighborhood $\Omega_R \backslash \Omega_{R-\varepsilon}$ of the edges $\partial\Gamma_R$. Since $\nabla v^R \in L^2(\Omega_R)$ and $v^R|_{\partial\Omega(R)} = 0$ in a neighborhood of the edges, we can apply Poincaré's inequality on the cross sections ω_j and integrate with respect to z_j to see $v^R \in L^2(\Omega_R)$. Since the Poiseuille flow $\mathcal{U}^1 \cdot (H + H_\infty)$ is smooth, i.e., $\mathcal{U}^1 \cdot (H + H_\infty) \in \mathcal{D}^l H(\Omega_R)$, we prove the regularity of $U = u^R - \mathcal{U}^1 \cdot (H + H_\infty)$ in $\Omega_R \backslash \Omega_{R-\varepsilon}$.

We fix a single outlet Q_j with the cross section ω_j, together with the local coordinates (y, z) and $\varepsilon < 1/2$. Then $U = \binom{V}{P}$ fulfills

$$\begin{aligned} SU = 0 \text{ in } \omega_j \times (R-\varepsilon, R), \quad V = 0 \text{ on } \partial\omega_j \times (R-\varepsilon, R), \\ FU = 0 \quad \text{for } z = R. \end{aligned} \tag{4.18}$$

We use the transformation $z \to \zeta = z - R$. If we observe the representation of V in local coordinates, i.e., $V = V_1 \mathrm{e}_1 + V_2 \mathrm{e}_2 + V_3 \mathrm{e}_z$ where e_1, e_2 are the unit vectors in y_1, y_2-direction, and e_z in the direction of the cylinder axis, respectively, then (4.18) is equivalent to

$$SU(y, \zeta) = 0 \quad \text{for } (y, \zeta) \in \omega_j \times (-\varepsilon, 0), \tag{4.19}$$

$$V(y, \zeta) = 0 \quad \text{for } (y, \zeta) \in \partial\omega_j \times (-\varepsilon, 0) \tag{4.20}$$

together with the boundary condition on the cross section in local coordinates:

$$V_3(y, 0) = 0, \tag{4.21}$$

$$\partial_1 V_3(y, 0) + \partial_z V_1(y, 0) = \partial_z V_1(y, 0) = 0,$$

$$\partial_2 V_3(y, 0) + \partial_z V_2(y, 0) = \partial_z V_2(y, 0) = 0. \tag{4.22}$$

Now we extend U on $Q_\varepsilon = \omega_j \times (-\varepsilon, \varepsilon)$, namely V_1, V_2, and P even, i.e., $U_i(y, \zeta) = U_i(y, -\zeta)$, $i = 1, 2, 4$, and V_3 odd, i.e., $V_3(y, \zeta) = -V_3(y, -\zeta)$. Then $V(y, \zeta) = 0$ for $y \in \partial\omega_j$, $|\zeta| < \varepsilon$. It is clear that $U \in L^2(Q_\varepsilon)^4$ and $S(U)$ exists as a distribution (on Q_ε), which is regular on $Q_\varepsilon^- = \omega_j \times (-\varepsilon, 0)$ and on $Q_\varepsilon^+ = \omega_j \times (0, \varepsilon)$. We see $SU = 0$ on Q_ε^- and on Q_ε^+ by elementary calculations. We show that no additional terms appear on the cross section $\omega_j \times \{0\}$. For this purpose we choose (a real valued) $\varphi \in C_0^\infty(Q_\varepsilon)$; we denote by $-\langle \operatorname{div} V, \varphi \rangle_{Q_\varepsilon}$ the application of the distribution $-\operatorname{div} V$ on φ. By definition,

$$-\langle \operatorname{div} V, \varphi \rangle_{Q_\varepsilon} = (V, \nabla\varphi)_{Q_\varepsilon} = (V, \nabla\varphi)_{Q_\varepsilon^-} + (V, \nabla\varphi)_{Q_\varepsilon^+}.$$

To each integral we apply the Gauss theorem. Observing that $\varphi = 0$ on ∂Q_ε and $\lim_{\zeta \to 0^+} V_3(y, \zeta) = \lim_{\zeta \to 0^-} V_3(y, \zeta) = 0$ in the trace sense, we obtain

$$\begin{aligned} -\langle \operatorname{div} V, \varphi \rangle_{Q_\varepsilon} &= -\int_{Q_\varepsilon^-} \operatorname{div} V \varphi \, \mathrm{d}x + \int_{\omega_j} V_3(y, 0) \varphi(y, 0) \, \mathrm{d}y \\ &- \int_{Q_\varepsilon^+} \operatorname{div} V \varphi \, \mathrm{d}x + \int_{\omega_j} \lim_{\zeta \to 0^+} V_3(y, \zeta) \varphi(y, \zeta) \, \mathrm{d}y = 0; \end{aligned}$$

hence div $V = 0$ on Q_ε in the distributional sense. Since div $V = 0$, we can write $Su = -\text{div } TU$ with $TU = \nabla U + (\nabla U)^\top - P\mathbb{I}$, where $\mathbb{I}$ is the unit matrix in $\mathbf{R}^3$. Now we repeat the same procedure for $-\Delta V + \nabla P = -\text{div } TU$ and $\Phi \in C_0^\infty(Q_\varepsilon)^3$. Since div $TU = 0$ in Q_ε^- and Q_ε^+, the trace of $TU \cdot \mathrm{e}_z = \lim_{\zeta \to 0} TU(y,\zeta) \cdot \mathrm{e}_z$ exists for $\zeta > 0$ and $\zeta < 0$. Hence we can calculate

$$\begin{aligned} -\langle \text{div } TU, \Phi\rangle_{Q_\varepsilon} &= (TU, \nabla\Phi)_{Q_\varepsilon} \\ &= -(\text{div } TU, \Phi)_{Q_\varepsilon^-} + \int_{\omega_j} TU(y,0) \cdot \mathrm{e}_z \cdot \Phi(y,0)\,\mathrm{d}y \\ &\quad -(\text{div } TU, \Phi)_{Q_\varepsilon^+} - \int_{\omega_j} \lim_{\zeta \to 0^+} TU(y,\zeta) \cdot \mathrm{e}_z \cdot \Phi(y,\zeta)\,\mathrm{d}y. \end{aligned} \tag{4.23}$$

From the definition of $U(y,\zeta)$ for $\zeta > 0$ we obtain for the first two components of $TU \cdot \mathrm{e}_z$

$$\partial_{y_i} V_3(y,\zeta) + \partial_z V_i(y,\zeta) = -\partial_{y_i} V_3(y,-\zeta) - \partial_z V_i(y,-\zeta) \to 0 \text{ with } \zeta \to 0, i = 1,2,$$

by (4.21) and (4.22). For the third component we get for $\zeta > 0$

$$\mathrm{e}_z^\top \cdot TU(y,\zeta) \cdot \mathrm{e}_z = 2\partial_z V_3(y,\zeta) - P(y,\zeta) = 2\partial_z V_3(y,-\zeta) - P(y,-\zeta);$$

hence

$$\lim_{\zeta \to 0^+} \mathrm{e}_z^\top \cdot TU(y,\zeta) \cdot \mathrm{e}_z = \lim_{\zeta \to 0^-} \mathrm{e}_z^\top \cdot TU(y,\zeta) \cdot \mathrm{e}_z.$$

Therefore the $\int_{\omega_j} \dots$ cancel each other in (4.23), which leads to $-\langle \text{div } TU, \Phi\rangle_{Q_\varepsilon} = 0$. Thus we obtain $SU = 0$ in Q_ε and $V = 0$ on $\partial\omega_j \times (-\varepsilon, \varepsilon)$. We apply the results of [48] and [6] once more and obtain $U \in \mathcal{D}^l H(Q_{\varepsilon'})$ for all $\varepsilon' < \varepsilon$, which finishes the proof. □

5. Uniform estimates for the solutions in Ω_R. To derive an error estimate for $u - u^R$, we need uniform estimates for the solutions of the system

$$Su = 0 \text{ in } \Omega_R, \quad v = 0 \text{ on } \partial\Omega(R), \quad Fu = h \text{ on } \Gamma_R \tag{5.1}$$

for $u \in \mathcal{D}_l^l V(\Omega_R)$. We denote the space of traces of $\{Fu : u \in \mathcal{D}_l^l V(\Omega_R)\}$ by $\mathcal{R}_l^l V(\Gamma_R, \partial\Gamma_R)$. The main result of this section is the following theorem.

THEOREM 5.1. *Let $l \in \mathbf{N}$ and $h \in \mathcal{R}_l^l V(\Gamma_R, \partial\Gamma_R)$ be given with*

$$\int_{\Gamma_R} \nu^\top \cdot h\,\mathrm{d}o = 0. \tag{5.2}$$

Let $u = (v,p) \in \mathcal{D}_l^l V(\Omega_R)$ be the unique solution of (5.1) with $\int_{G_0} p\,\mathrm{d}x = 0$, where G_0 is chosen as in Theorem 4.4. Then the following holds:

$$\|u; \mathcal{D}_l^l V(\Omega_R)\| \le C\,R^2\,\|h; \mathcal{R}_l^l V(\Gamma_R, \partial\Gamma_R)\|, \tag{5.3}$$

where C is a constant independent of R.

Main idea of the proof. The main idea of the proof is the following: Let $X = \{u \in \mathcal{D}_l^l V(\Omega_R), Su = 0 \text{ in } \Omega_R, v = 0 \text{ on } \partial\Omega(R), \int_{G_0} p\,\mathrm{d}x = 0\}$. Then X is a closed linear subspace of $\mathcal{D}_l^l V(\Omega_R)$, and by Theorem 4.2, the operator

$$F : u \to Fu \in Y := \left\{ h \in \mathcal{R}_l^l V(\Gamma_R), \int_{G_R} \nu^\top \cdot h\,\mathrm{d}o = 0 \right\}$$

defines an isomorphism. To prove (5.3), we construct an "almost inverse operator," i.e., a continuous linear operator $\mathcal{A} : Y \to X$, such that $F\mathcal{A} = \mathbb{I} + \mathcal{F}$, where $\mathbb{I}$ is the identity on Y, $\mathcal{F}$ is small for large R, i.e., there exists $q < 1$ and $R_0 > 0$, such that

$$\|\mathcal{F} : Y \to Y\| < q \tag{5.4}$$

for all $R > R_0$. Moreover, we prove

$$\|\mathcal{A} : Y \to X\| \le CR^2, \tag{5.5}$$

where C is independent of R. Then $(\mathbb{I} + \mathcal{F})^{-1}$ exists as a Neumann series, and we obtain $F^{-1} = \mathcal{A}(\mathbb{I} + \mathcal{F})^{-1}$. Now (5.4) and (5.5) lead to (5.3).

For the construction of $\mathcal{A}$ we need existence and uniqueness results for the *second limit problem* in suitable function spaces. The second limit problem is the problem on the semicylinder formulated in the following form. Let $\omega \subset \mathbf{R}^2$ be a bounded domain with $\partial\omega \in C^{l+2}$. We set

$$\Pi^- = \omega \times (-\infty, 0), \quad x = (y, z) \text{ for } x \in \Pi^-, \quad \partial\Pi_{(-)} = \partial\omega \times (-\infty, 0).$$

By $\omega(z)$ and $\partial\omega(z)$ we denote the cross section and its boundary in $z \le 0$; in this notation $\partial\omega(0)$ is the edge of the semicylinder. On Π^- we consider the following boundary value problem:

$$Su = f \text{ in } \Pi^-, \quad v = g_0 \text{ on } \partial\Pi_{(-)}, \tag{5.6}$$

$$Fu = g_1 \text{ on } \omega \times \{0\}, \quad (Fu)^\top = (\mathrm{e}_1^\top \cdot Tu \cdot \mathrm{e}_z, \mathrm{e}_2^\top \cdot Tu \cdot \mathrm{e}_z, v_3). \tag{5.7}$$

As before, e_i denote the unit vectors in y_i-direction for $i = 1, 2$. The boundary of Π^- has two types of singularities: the edge $M = \partial\omega \times \{0\}$ and the cylindrical outlet to infinity. To treat the full asymptotic behavior in the vicinity of the edges and for $z \to -\infty$ we must use weighted spaces with different weights near $\partial\omega(0)$ and at $-\infty$.

We choose basic weight functions $\rho_1, \rho_2 \in C^{l+2}(\Pi^-)$ with the following properties: We define ρ_1 analogous to the weight ρ in section 3.2. $\rho_1(x) = \operatorname{dist}(x, \partial\omega(0))$ for $x \in \mathcal{O}_\varepsilon$, $\rho_1(y, z) = 1$ for $z < -2\varepsilon$, where ε is chosen small enough such that $\operatorname{dist}(x, \partial\omega(0))$ is a C^{l+2}-function on $\mathcal{O}_\varepsilon \cap \Pi^- = \{x \in \Pi^- : \operatorname{dist}(x, \partial\omega(0)) < \varepsilon\}$. We set $\rho_2(y, z) = \rho_2(z) = e^z$ for $z < -2$, $\rho_2(z) = 1$ for $-1 < z \le 0$. For $\gamma, \beta \in \mathbf{R}$, $l \in \mathbf{N}$, $\varphi \in C_0^\infty(\overline{\Pi}^- \setminus \partial\omega(0))$ we introduce the norm

$$\|\varphi; \mathbb{W}^l_{\gamma,\beta}(\Pi^-)\| = \sum_{|\alpha| \le l} \left(\|(\rho_1^{\gamma - l + |\alpha|} \rho_2^\beta \partial^\alpha \varphi; L^2(\Pi^-)\|^2 \right)^{1/2} \tag{5.8}$$

and $\mathbb{W}^l_{\gamma,\beta}(\Pi^-)$ as the closure of $C_0^\infty(\overline{\Pi}^- \setminus \partial\omega(0))$ in the norm (5.8). $\mathbb{W}^l_{\gamma,\beta}(\Pi^-)$ coincides with the space of all $\varphi \in H^l_{loc}(\Pi^-)$ such that $\|\varphi; \mathbb{W}^l_{\gamma,\beta}(\Pi^-)\| < \infty$. For $0 \le k \le l$, $\delta, \delta' \ge 0$, the space $\mathbb{W}^l_{\gamma,\beta}(\Pi^-)$ is continuously embedded into $\mathbb{W}^{l-k}_{\gamma-k+\delta,\beta+\delta'}(\Pi^-)$. For $k > 0$, $\delta, \delta' > 0$ this embedding is compact (see [52] for details).

The spaces of traces are defined in a natural way. Again we divide them into trace spaces on the lateral surface $\partial\Pi_{(-)}$ and on the cross section $\omega(0)$. We set $\mathbb{W}^{l-1/2}_{\gamma,\beta}(\partial\Pi_{(-)}) = \{\varphi|_{\partial\Pi_{(-)}} : \varphi \in \mathbb{W}^l_{\gamma,\beta}(\Pi^-)\}$ provided with the canonical norm $\|\varphi; \mathbb{W}^{l-1/2}_{\gamma,\beta}(\partial\Pi_{(-)})\| = \inf \|\tilde{\varphi}; \mathbb{W}^l_{\gamma,\beta}(\Pi^-)\|$, where the infimum is taken over all functions $\tilde{\varphi}$ with $\tilde{\varphi}|_{\partial\Pi_{(-)}} = \varphi$. On the cross section $\omega(0)$ the space of traces of $\mathbb{W}^l_{\gamma,\beta}(\Pi^-)$

coincides with $V_\gamma^{l-1/2}(\omega,\partial\omega)$, the trace space of $V_\gamma^l(\Pi^-,\partial\omega(0)) = \mathbb{W}_{\gamma,0}^l(\Pi^-)$ (where the exponential weight is equal to 1 everywhere). We define the natural domain and range of the problem (5.6), (5.7):

$$\begin{aligned}
\mathcal{D}_{\gamma,\beta}^l\mathbb{W}(\Pi^-) &= \mathbb{W}_{\gamma,\beta}^{l+1}(\Pi^-)^3 \times \mathbb{W}_{\gamma,\beta}^l(\Pi^-),\\
\mathcal{R}_{\gamma,\beta}^l\mathbb{W}(\Pi^-) &= \mathbb{W}_{\gamma,\beta}^{l-1}(\Pi^-)^3 \times \mathbb{W}_{\gamma,\beta}^l(\Pi^-),\\
\mathcal{R}_{\gamma,\beta}^l\mathbb{W}(\Pi^-,\partial\Pi_{(-)},\omega(0)) &= \mathcal{R}_{\gamma,\beta}^l\mathbb{W}(\Pi^-) \times \mathbb{W}_{\gamma,\beta}^{l+1/2}(\partial\Pi_{(-)})^3\\
&\quad \times V_\gamma^{l-1/2}(\omega(0),\partial\omega)^2 \times V_\gamma^{l+1/2}(\omega(0),\partial\omega).
\end{aligned}$$

Then the operator

$$\begin{aligned}
\mathbb{S}_{\gamma,\beta}^l : \mathcal{D}_{\gamma,\beta}^l\mathbb{W}(\Pi^-) &\to \mathcal{R}_{\gamma,\beta}^l\mathbb{W}(\Pi^-,\partial\Pi_{(-)},\omega(0))\\
u &\to (Su, Du|_{\partial\Pi_{(-)}}, Fu|_{\omega(0)})
\end{aligned} \tag{5.9}$$

defines a continuous linear operator. From the results of the previous sections it is evident for which exponents (5.9) can be expected to be Fredholm, namely, for $|\gamma - l| < 1$ and for all β such that the line $\operatorname{Im}\lambda = \beta$ is free of eigenvalues of the elliptic pencil belonging to the Dirichlet problem of the Stokes system in the cylinder $\Pi = \omega \times \mathbf{R}$, where ω is one of the cross sections ω_j. We formulate the special results we need in the following lemma.

LEMMA 5.2. *If $0 < \beta < \beta_*$, the Stokes problem* (5.6), (5.7) *has a solution $u \in \mathcal{D}_{l,\beta}^l\mathbb{W}(\Pi^-)$ for all $(f,g,h) \in \mathcal{R}_{l,-\beta}^l\mathbb{W}(\Pi^-,\partial\Pi_{(-)},\omega(0))$. This solution is uniquely determined up to a constant in the pressure and admits the asymptotic representation*

$$u = -Hu^{j1} + au^{\#} + \tilde{u}, \tag{5.10}$$

where $\tilde{u} \in \mathcal{D}_{l,-\beta}^l\mathbb{W}(\Pi^-)$ and

$$H = \int_{\Pi^-} f_4\,\mathrm{d}x + \int_{\partial\Pi_{(-)}} \nu^\top \cdot g\,\mathrm{d}o + \int_\omega h_3\,\mathrm{d}y.$$

The constant in pressure can be fixed, e.g., by the condition $\int_\omega p(y,-1)\,\mathrm{d}y = 0$. In this case the following estimate holds

$$|a| + \|\tilde{u}; \mathcal{D}_{l,-\beta}^l\mathbb{W}(\Pi^-)\| \le C\|(f,g,h); \mathcal{R}_{l,-\beta}^l\mathbb{W}(\Pi^-,\partial\Pi^-)\|. \tag{5.11}$$

Proof. Combining the results on the Stokes problem (1.1) in the straight cylinder $\Pi = \omega\times\mathbf{R}$ with the results in domains with smooth edges gives the Fredholm property of the mapping (5.9) if $|\gamma - l| < 1$ and the line $\operatorname{Im}\lambda = \beta$ is free of eigenvalues of the pencils $\mathfrak{S}(\lambda)$ associated to the Dirichlet problem (1.1) in the cylinder Π (see [33, Theorem 3.1.1, Theorem 8.1.1, and section 4.1.2]). From the results cited above, we obtain $\operatorname{Ind}\mathbb{S}_{l,\beta}^l - \operatorname{Ind}\mathbb{S}_{l,-\beta}^l = 2$ for $0 < \beta < \beta^*$. Moreover, if $(u^*,g^*,h^*) \in \ker\,(\mathbb{S}_{l,\beta}^l)^*$ (= kernel of the adjoint operator), then $u^* \in \ker\mathbb{S}_{l,-\beta}^l(\Pi^-)$, while $g^* = N_1u^*$ on $\partial\Pi_{(-)}$, $h^* = F_0u^*$ on $\omega(0)$.

If $0 < \beta < \beta^*$, then the operator $\mathbb{S}_{l,-\beta}^l$ is injective. Indeed, let $u \in \mathcal{D}_{l,-\beta}^l\mathbb{W}(\Pi^-)$ be a solution to the homogeneous problem (5.6), then $u \in H^{l+1}(\Pi^-)^3 \times H^l(\Pi^-)$. For $z \to -\infty$ this follows from the fact that the weight increases exponentially, in the neighborhood of the edge it follows from the mirror principle, as in the proof

of Theorem 4.4. Since $Su = 0$ iff $S_1 u = 0$ we can multiply the equation $S_1 u = 0$ scalar with u (in $L^2(\Pi^-)^4$) and integrate by parts. This gives $\nabla v + (\nabla v)^\top = 0$; hence, $v = A \times x + B$ with constant vectors A, $B \in \mathbb{C}$. The homogeneous boundary conditions lead to $u = 0$ then. Applying the results mentioned above gives the surjectivity of $\mathbb{S}^l_{l,\beta}$. Now elementary calulations together with the index formula lead to

$$\dim \ker \mathbb{S}^l_{l,\beta} = \dim \operatorname{coker} \mathbb{S}^l_{l,-\beta} = 1.$$

Then it is clear that $\ker \mathbb{S}^l_{l,\beta} = \{cu^\# : c \in \mathbb{C}\}$.

Now let $u \in \mathcal{D}^l_{l,\beta}\mathbb{W}(\Pi^-)$ be any solution to (5.6), (5.7) for given f, g, and h as above. Formula (1.21) (now for $J = 1$) ensures the existence of constants H and a, such that (5.10) holds. The cut-off function may be omitted in this case, since by Lemma 4.6, $u^{j1} \in \mathcal{D}^l_l V(\omega \times (-c_0, 0), \partial\omega(0))$ for any $c_0 < 0$, for $u^\#$ this is elementary. We apply the generalized Green's formula (1.24) to u and $U = u^\#$ and obtain (observing that $\pi_0(\,\cdot\,)$, $\pi_1(\,\cdot\,)$ consist of one component now)

$$\int_{\Pi^-} f_4 \,\mathrm{d}x + \int_{\partial\Pi_{(-)}} \nu^\top \cdot g \,\mathrm{d}o + \int_\omega h_3 \,\mathrm{d}y = \langle \pi_0 u, \pi_1 u^\# \rangle - \langle \pi_1 u, \pi_0 u^\# \rangle = H. \qquad \square$$

Proof of Theorem 5.1. We recall that $\Gamma_{R,j}$ denotes the cross section at $z_j = R$ in the outlet Q_j; we set $h_{(j)} = h|_{\Gamma_{R,j}}$. We define $H_j = \int_{\Gamma_{R,j}} \nu^\top \cdot h \,\mathrm{d}o$, this means in local coordinates of Q_j: $H_j = \int_{\omega_j} h_{(j),3}(y) \,\mathrm{d}y$, and we obtain

$$|H_j| \le C\|h; L^2(\Gamma_R)\| \le C\|h; \mathcal{R}^l_l V(\Gamma_R, \partial\Gamma_R)\|. \tag{5.12}$$

Now we fix β with $\beta < \beta^*$ arbitrary. Let $u_{(j)} \in \mathcal{D}^l_{l,\beta}(\Pi^-_j)$ be the unique solution of the problem

$$Su_{(j)} = 0 \text{ in } \Pi^-_j = \omega_j \times (-\infty, 0), \quad v_{(j)} = 0 \text{ on } \partial\omega_j \times (-\infty, 0), \tag{5.13}$$

$$Fu_{(j)} = h_{(j)} \text{ on } \omega_j(0), \quad \int_{\omega_j} p_{(j)}(y, -1) = 0. \tag{5.14}$$

According to (5.10), $u_{(j)}$ admits the representation $u_{(j)} = -H_j u^{j1} + a_j u^\# + \tilde{u}_{(j)}$ where, as before, u^{j1} is the Poiseuille flow in the cylinder Π_j and $u^\# = (0, 0, 0, 1)$ the constant pressure solution. The following estimate holds:

$$|a_j| + \|\tilde{u}_{(j)}; \mathcal{D}^l_{l,-\beta}\mathbb{W}(\Pi^-_j)\| \le C\|h_{(j)}; \mathcal{R}^l_l V(\omega_j, \partial\omega_j)\|. \tag{5.15}$$

Let χ_j denote the cut-off functions of section 1. For $x \in Q_j \cap \Omega_R$ the expression $\chi_j(z_j) u_{(j)}(y, z_j - R)$ is well defined and can be extended smoothly by 0 to the whole domain Ω_R. In this sense we set

$$u_\Pi = \sum_{j=1}^J \chi_j(z_j) u_{(j)}(y_j, z_j - R). \tag{5.16}$$

Recalling the notations of section 1, i.e.,

$$a = (a_1, \dots, a_J), \quad H = (H_1, \dots, H_J), \quad \mathcal{U}^h = (\chi_1 u^{1h}, \dots, \chi_J u^{Jh}),$$

$h = 0, 1$, where u^{j0} is the constant pressure solution and u^{j1} the Poiseuille flow in the outlet Q_j, then, with $\varpi = (\varpi_1, \dots, \varpi_J)$,

$$u_\Pi = -\mathcal{U}^1 \cdot H|_{\Omega_R} + \mathcal{U}^0 \cdot (a + 2R\, H\varpi)|_{\Omega_R} + \sum_{j=1}^J \chi_j(z_j) \tilde{u}_{(j)}(y_j, z_j - R). \tag{5.17}$$

The term $2R\,H\varpi = 2R(H_1\varpi_1, \dots, H_J\varpi_J)$ appears from the special structure of the Poiseuille flow. Then Su_Π has a compact support contained in $\bigcup_{j=1}^J \operatorname{supp} \nabla\chi_j \subset \Omega_0$; moreover, $v_\Pi = 0$ on $\partial\Omega(R)$ and $p_\Pi = 0$ on G_0 since $\operatorname{supp}\chi_j \cap G_0 = \emptyset$ for $j = 1, \dots, J$. We extend Su_Π smoothly by 0 to $f_\infty \in \mathcal{R}^l_\beta W(\Omega)$. The norm of f_∞ can be estimated by

$$\begin{aligned}\|f_\infty; \mathcal{R}^l_\beta W(\Omega)\| &\le C\|f_\infty; R^l H(\Omega_0)\| \\ &\le C\Big(|H|\,\|\mathcal{U}^1; \mathcal{D}^l H(\Omega_0)\| + |a + 2H\varpi R|\,\|\mathcal{U}^0; \mathcal{D}^l H(\Omega_0)\| \\ &\qquad + \sum_{j=1}^J \|\tilde{u}_{(j)}; \mathcal{D}^l H(\omega_j \times (-R, -R + R_0))\|\Big).\end{aligned}$$

The last term can be estimated by $e^{-\beta(R+R_0)}\|\tilde{u}_{(j)}; \mathcal{D}^l_{l,-\beta}\mathbb{W}(\Pi_j^-)\|$, thus decays exponentially, and with (5.12) and (5.15) it follows that

$$\|f_\infty; \mathcal{R}^l_\beta W(\Omega)\| \le CR\|h; Y\|. \tag{5.18}$$

Furthermore, we have $\int_\Omega f_{\infty,4}\,\mathrm{d}x = 0$: Since $v_\Pi = 0$ on $\partial\Omega(R)$ and $Fu_\Pi = h$ on Γ_R by construction, then Gauss' theorem gives, with (5.2),

$$\int_\Omega f_{\infty,4}\,\mathrm{d}x = -\int_{\Omega_R} \operatorname{div} u_\Pi\,\mathrm{d}x = -\int_{\partial\Omega_R} \nu^\top \cdot v_\Pi\,\mathrm{d}o = -\int_{\Gamma_R} \nu^\top \cdot h\,\mathrm{d}o = 0.$$

Thus, by Theorem 1.6, we obtain a unique solution $u_\infty \in \mathbb{D}^l_{\pm\beta}W(\Omega)$ of the problem $Su_\infty = f_\infty$ in Ω, $v_\infty = 0$ on $\partial\Omega$, $\pi_1 u_\infty = 0$, $\int_{G_0} p_\infty = 0$. u_∞ admits the representation

$$u_\infty = \mathcal{U}^0 \cdot a^* + \tilde{u}_\infty, \tag{5.19}$$

where

$$|a^*| + \|\tilde{u}_\infty; \mathcal{D}^l_\beta W(\Omega)\| \le C\|f_\infty; \mathcal{R}^l_\beta W(\Omega)\|. \tag{5.20}$$

Now we define

$$U = (V, P) = \mathcal{A}(h) = u_\Pi + u_\infty.$$

By construction it is clear that $SU = 0$ in Ω_R, $V = 0$ on $\partial\Omega(R)$, and $\int_{G_0} P\,\mathrm{d}x = 0$. Moreover, from Lemma 4.6 we have $u_\infty|_{\Omega_R} \in \mathcal{D}^l_l V(\Omega_R, \partial\Gamma_R)$; hence, $U \in X$ and $FU \in \mathcal{R}^l_l V(\Gamma_R, \partial\Gamma_R)$. As already mentioned in section 1, the solution u_∞ carries no flux through the outlets, which means

$$\int_{\Gamma_{R,j}} \nu^\top \cdot v_\infty\,\mathrm{d}o = 0$$

for $j = 1, \dots, J$; hence $\int_{\Gamma_R} \nu^\top \cdot FU\,\mathrm{d}o = 0$, which ensures $FU \in Y$. Since $F(\mathcal{U}^0 \cdot a^*) = 0$ on Γ_R, it follows $FU = Fu_\Pi + F\tilde{u}_\infty = h + \mathcal{F}(h)$.

Let $\mathcal{O}_\epsilon = \{x \in \Omega_R, \operatorname{dist}(x, \Gamma_R) < \epsilon\}$ be a neighborhood of Γ_R of constant size; then the trace theorem, together with (5.18) and (5.20), gives

$$\begin{aligned}\|\mathcal{F}(h); Y\| &\le \|\tilde{u}_\infty; \mathcal{D}^l_l V(\mathcal{O}_\epsilon, \partial\Gamma_R)\| \\ &\le Ce^{-\beta R}\|f_\infty; \mathcal{R}^l_\beta W(\Omega)\| \le Ce^{-\beta R}\,R\,\|h; Y\|,\end{aligned}$$

where C is a constant independent of R. Here we have to observe that by the definition of the norms in $W^l_\beta(\Omega)$ and Lemma 4.6

$$\|\tilde{u}_\infty; \mathcal{D}^l_l V(\mathcal{O}_\epsilon, \partial\Gamma_R)\| \le C\|\tilde{u}_\infty; \mathcal{D}^l H(\mathcal{O}_\epsilon)\| \le Ce^{-\beta R}\|\tilde{u}_\infty; \mathcal{D}^l_\beta W(\Omega)\|.$$

We remember that β can be chosen arbitrarily in the interval $(0, \beta^*)$. Now we fix R_* large enough, such that $Ce^{-\beta R_*}R_* \le q < 1$. Then for all $R > R_*$, it holds $\|\mathcal{F} : Y \to Y\| \le q$, and

$$(\mathbb{I} + \mathcal{F})^{-1} = \sum_{k=0}^{\infty}(-1)^k \mathcal{F}^k, \quad \|(\mathbb{I} + \mathcal{F})^{-1}\| \le \frac{1}{1+q}. \tag{5.21}$$

To estimate U, we first estimate u_Π with the help of the representation (5.17), estimate (5.15), and $E = (1, \dots, 1)^\top$:

$$\begin{aligned}
&\|u_\Pi; \mathcal{D}^l_l V(\Omega_R)\| \\
&\le C\Bigg(|H|\|\mathcal{U}^1 \cdot E; \mathcal{D}^l H(\Omega_R)\| + |a + 2H\varpi R|\|\mathcal{U}^0 \cdot E; \mathcal{D}^l H(\Omega_R)\| \\
&\qquad + \sum_{j=1}^{J} \|\tilde{u}_{(j)}; \mathcal{D}^l_l V(\omega_j \times (-R, 0))\|\Bigg) \\
&\le C\left(R^2|H| + R|a + 2H\varpi R| + \sum_{j=1}^{J} \|\tilde{u}_{(j)}; \mathcal{D}^l_{l,-\beta}\mathbb{W}(\Pi^-_j)\|\right);
\end{aligned}$$

hence

$$\|u_\Pi; \mathcal{D}^l_l V(\Omega_R)\| \le CR^2\|h; Y\|. \tag{5.22}$$

For u_∞ we calculate by means of (5.19), Lemma 4.6, (5.20), and (5.18):

$$\begin{aligned}
\|u_\infty; \mathcal{D}^l_l V(\Omega_R)\| &\le C\Big(|a^*|\,\|\mathcal{U}^0 \cdot E; \mathcal{D}^l_l V(\Omega_R)\| + \|\tilde{u}_\infty; \mathcal{D}^l_l V(\Omega_R)\|\Big) \\
&\le C\Big(|a^*|\,R\,\|f_\infty; \mathcal{R}^l_\beta W(\Omega)\|\Big) \\
&\le C\,R\,\|f_\infty; \mathcal{R}^l_\beta W(\Omega)\| \le C\,R^2\,\|h; Y\|.
\end{aligned} \tag{5.23}$$

Inequalities (5.21)–(5.23) lead to (5.3), and the theorem is proved. □

For the approximation of weak solutions with $L^2(\Omega)$-forces we need the following corollary.

COROLLARY 5.3. *Let $l \in \mathbf{N}$ and $h \in \mathcal{R}^l_l V(\Gamma_R, \partial\Gamma_R)$ be given with*

$$\int_{\Gamma_{R,j}} \nu^\top \cdot h \,\mathrm{d}o = 0 \quad \text{for } j = 1, \dots, J. \tag{5.24}$$

Let $u = \binom{v}{p} \in \mathcal{D}^l_l V(\Omega_R)$ be the unique solution of (5.1) with $\int_{G_0} p\,\mathrm{d}x = 0$. Then it holds that

$$\|v; V^{l+1}_l(\Omega_R, \partial\Gamma_R)^3\| + \|\nabla p, V^{l-1}_l(\Omega_R, \partial\Gamma_R)^3\| \le C\|h; \mathcal{R}^l_l V(\Gamma_R, \partial\Gamma_R)\|, \tag{5.25}$$

where C is a constant independent of R.

Proof. The corollary reflects the special case where $H_j = 0$, $j = 1, \dots, J$ in the proof of Theorem 5.1. Thus in the estimates of f_∞ and u_Π, all terms with H vanish, and we obtain

$$\|f_\infty; \mathcal{R}_l^l V(\Omega)\| \le C\|h; Y\| \tag{5.26}$$

instead of (5.18). If we observe that all terms with $\mathcal{U}^0$ influence only the estimates for the pressure; and moreover, $\nabla \mathcal{U}^0 = 0$ for $z_j > R_0$, $j = 1, \dots, J$, then it holds that

$$\|v_\Pi; V_l^{l+1}(\Omega_R, \partial\Gamma_R)\| + \|\nabla p_\Pi, V_l^{l-1}(\Omega, \partial\Gamma_R)\| \le C\|h; Y\| \tag{5.27}$$

instead of (5.22). Similarly, for u_∞ we obtain

$$\|v_\infty; V_l^{l+1}(\Omega, \partial\Gamma_R)^3\| + \|\nabla p_\infty; V_l^{l-1}(\Omega_R, \partial\Gamma_R)^3\| \le C\|h; Y\| \tag{5.28}$$

instead of (5.23). (5.25)–(5.28) lead to (5.24), which proves the assertion. □

6. The error estimate. With the result of the previous section we are now able to prove the error estimate for the approximation problem as it is defined in section 2. We recall the definition of the cut-off function χ_R. We assume that $\{\chi_R\}_R$ is a system of C^∞-functions in Ω with the properties $\chi_R \equiv 1$ on Ω_0; in Q_j we require $\chi_R(x) = \chi_R(z_j) = 1$ for $z_j < R - 1$, $\chi_R(z_j) = 0$ for $z_j > R - 1/2$.

THEOREM 6.1. *Let $l \in \mathbf{N}$, $\beta^* > \beta > 0$ meet the requirements of Theorem* 1.6, *$G_0 \subset \Omega_0$ be a nonvoid subdomain such that $G_o \cap \operatorname{supp} \chi_j = \emptyset$ for all j, and let $(f, g) \in \mathcal{R}_\beta^l W(\Omega, \partial\Omega)$, $H \in \mathbb{C}^J$ be given such that the flux condition*

$$\int_\Omega f_4 \,\mathrm{d}x + \int_{\partial\Omega} \nu^\top \cdot g \,\mathrm{d}o + \sum_{j=1}^J H_j = 0 \tag{6.1}$$

is fulfilled. Let $u \in \mathbb{D}_{\pm\beta}^l W(\Omega)$ be the unique solution of the Stokes problem with prescribed fluxes according to Theorem 1.6 *and $u^R \in \mathcal{D}^l H(\Omega_R)$ be the unique solution to the approximation problem*

$$\begin{aligned} Su^R = \chi_R f \text{ in } \Omega_R, \quad v^R = \chi_R g \text{ on } \partial\Omega(R), \\ Fu^R = F(\mathcal{U}^1 \cdot (H + H_\infty)) \text{ on } \Gamma_R, \quad \int_{G_0} p^R = 0, \end{aligned} \tag{6.2}$$

where $H_\infty \in \mathbb{C}^J$ with

$$H_{\infty,j} = \int_{Q_j} (1 - \chi_R) f_4 \,\mathrm{d}x + \int_{\partial Q_j} \nu^\top \cdot (1 - \chi_R) g \,\mathrm{d}o. \tag{6.3}$$

Then for every $\varepsilon > 0$ we obtain the error estimate:

$$\begin{aligned} \|u - u^R; \mathcal{D}^l H(\Omega_{R-1})\| &= \|u - u^R; H^{l+1}(\Omega_{R-1})^3 \times H^l(\Omega_{R-1})\| \\ &\le \|u - u^R; \mathcal{D}_l^l V(\Omega_R, \partial\Gamma_R)\| \\ &\le C\, e^{-(\beta-\varepsilon)R} \Big(|H| + \|(f, g); \mathcal{R}_\beta^l W(\Omega, \partial\Omega)\| \Big), \end{aligned} \tag{6.4}$$

where C is a constant independent of R and of the data.

Proof. From the flux condition (6.1) and the definition of H_∞ we obtain

$$\int_{\Omega_R} \chi_R f_4 \,\mathrm{d}x + \int_{\partial\Omega(R)} \nu^\top \cdot \chi_R g \,\mathrm{d}o + \int_{\Gamma_R} \nu^\top \cdot \mathcal{U}^1 \cdot (H + H_\infty) \,\mathrm{d}o = 0;$$

hence $u^R \in \mathcal{D}^l H(\Omega_R)$ exists due to Theorem 4.4.

According to (1.22), u has the representation

$$u = \mathcal{U}^1 \cdot H + \mathcal{U}^0 \cdot a + \tilde{u}, \tag{6.5}$$

with $\tilde{u} \in \mathcal{D}^l_\beta W(\Omega)$. Let u_∞ be the solution of

$$\begin{aligned} &Su_\infty = \chi_R f \text{ in } \Omega, \quad v_\infty = \chi_R g \text{ on } \partial\Omega, \\ &\pi_1 u_\infty = H + H_\infty, \quad \int_{G_0} p_\infty = 0. \end{aligned} \tag{6.6}$$

Again, with (1.22), we have the following representation for u_∞:

$$u_\infty = \mathcal{U}^1 \cdot (H + H_\infty) + \mathcal{U}^0 \cdot a_\infty + \tilde{u}_\infty. \tag{6.7}$$

The difference $U = u - u_\infty$ is the unique solution of the problem

$$SU = (1-\chi_R)f \text{ in } \Omega, \quad DU = (1-\chi_R)g \text{ on } \partial\Omega, \quad \pi_1 U = -H_\infty, \quad \int_{G_0} P \,\mathrm{d}x = 0.$$

Since $\operatorname{supp}(1-\chi_R)(f,g) \subset \{x \in \Omega : z_j > R-1, j = 1,\dots,J\}$, for every $\bar{\varepsilon}$ with $\beta > \bar{\varepsilon} > 0$, estimate (1.29) leads to

$$\begin{aligned} &|a - a_\infty| + \|\tilde{u} - \tilde{u}_\infty; \mathcal{D}^l_{\bar{\varepsilon}} W(\Omega)\| \\ &\le C\Big(\|(1-\chi_R)(f,g); \mathcal{R}^l_{\bar{\varepsilon}} W(\Omega, \partial\Omega)\| + |H_\infty|\Big) \\ &\le C\Big(e^{-(\beta-\bar{\varepsilon})R} \|(1-\chi_R)(f,g); \mathcal{R}^l_\beta W(\Omega, \partial\Omega)\| + |H_\infty|\Big). \end{aligned} \tag{6.8}$$

To estimate $|H_\infty|$ we apply the Cauchy–Schwarz inequality to

$$\begin{aligned} |H_{\infty,j}| &\le |\int_{Q_j} (1-\chi_R) f_4 \,\mathrm{d}x| + |\int_{\partial Q_j} \nu^\top \cdot (1-\chi_R) g \,\mathrm{d}o| \\ &\le \Big(\int_{Q_j} |1-\chi_R|^2 e^{-2\beta z} \,\mathrm{d}x\Big)^{1/2} \Big(\|f_4 e^{\beta z}; L^2(Q_j)\| + \|g e^{\beta z}; L^2(\partial\Omega)\|\Big) \\ &\le C\, e^{-\beta R} (\|(f,g); \mathcal{R}^l_\beta W(\Omega, \partial\Omega)\|. \end{aligned} \tag{6.9}$$

(6.8) and (6.9), together with Lemma 4.6, lead to

$$\begin{aligned} &\|U; \mathcal{D}^l H(\Omega_{R-1})\| \le \|U; \mathcal{D}^l_l V(\Omega_R)\| \\ &\le C(R^2 |H_\infty| + R|a - a_\infty| + \|\tilde{u} - \tilde{u}_\infty; \mathcal{D}^l_{\bar{\varepsilon}} W(\Omega)\|) \\ &\le C\, e^{-(\beta-\varepsilon)R} \|(f,g); \mathcal{R}^l_\beta W(\Omega, \partial\Omega)\|, \end{aligned} \tag{6.10}$$

where ε in dependence of $\bar{\varepsilon}$ can be chosen arbitrarily small. Now we rewrite the difference $u - u^R$ in Ω_R: $u - u^R = u - u_\infty + u_\infty - u^R$. Then $U^R = u_\infty - u^R \in \mathcal{D}^l_l V(\Omega_R, \partial\Gamma_R)$. Moreover, due to (6.2) and (6.7), U^R solves the problem

$$(SU^R, DU^R) = (0,0) \text{ in } (\Omega_R, \partial\Omega(R)), \text{ and } FU^R = F\tilde{u}_\infty \text{ on } \Gamma_R, \int_{G_0} P^R \,\mathrm{d}x = 0.$$

We apply Theorem 5.1 and gain

$$\begin{aligned}\|U^R;\mathcal{D}^l H(\Omega_{R-1})\| &\le C\|U^R;\mathcal{D}^l_l V(\Omega_R)\|\\ &\le CR^2\|F\tilde{u}_\infty;\mathcal{R}^l_l V(\Gamma_R,\partial\Gamma_R)\|\\ &\le CR^2\|\tilde{u}_\infty;\mathcal{D}^l_l V(\mathcal{O}_\epsilon(\Gamma_R),\partial\Gamma_R)\| \le Ce^{-\beta R}R^2\|\tilde{u}_\infty;\mathcal{D}^l_\beta W(\Omega)\|\\ &\le Ce^{-\beta R}R^2\Big(|H+H_\infty|+\|\chi_R(f,g);\mathcal{R}^l_\beta W(\Omega,\partial\Omega)\|\Big)\\ &\le C\,e^{-(\beta-\varepsilon)R}\Big(|H|+\|(f,g);\mathcal{R}^l_\beta W(\Omega,\partial\Omega)\|\Big),\end{aligned}$$

which, together with (6.10), proves the theorem. □

Remark 6.2. We outline that estimate (6.4) does *not* imply that $\|u^R;H^{l+1}(\Omega_R)^3\times H^l(\Omega_R)\|$ remains bounded as $R\to\infty$, of course.

The estimate (6.4) can also be proved with small modifications for $\|u-u^R;\mathcal{D}^l_{l,\bar\beta}\mathbb{W}(\Omega_R,\partial\Gamma_R)\|$ with $0<\bar\beta<\beta$. Here the $\mathcal{D}^l_{l,\bar\beta}V(\Omega_R,\partial\Gamma_R)$ means that the weight $\rho(x)$ for spaces $V^l_l(\Omega_R,\partial\Gamma_R)$ is multiplied by $e^{\bar\beta\mathfrak{z}}$. In this case we obtain

$$\|u-u^R;\mathcal{D}^l_{l,\bar\beta}V(\Omega_R,\partial\Gamma_R)\|\le Ce^{-(\beta-\bar\beta-\varepsilon)R}\Big(\|(f,g);\mathcal{R}^l_\beta W(\Omega,\partial\Omega)\|+|H|\Big).$$

7. Approximation for L^2-data. As already mentioned, it is possible to treat the problem

$$-\Delta v+\nabla p=f,\quad \operatorname{div} v=0\ \text{ in }\Omega,\quad v=0\text{ on }\partial\Omega \tag{7.1}$$

within the theory of H^l spaces, at least for v. Since p is unique up to a constant, we only obtain estimates for ∇p, of course. The following result is well known (see [40, p. 84]).

THEOREM 7.1 (weak solutions with zero fluxes). *Let $\Omega\subset\mathbf{R}^3$ be a domain with J cylindrical outlets as in the previous chapter and with $\partial\Omega$ of class C^{l+2} for some $l\in\mathbf{N}$. For every $f\in H^{l-1}(\Omega)^3$ there exists a unique solution $(v,\nabla p)$ of (7.1) with $\nabla v\in L^2(\Omega)$. The following estimate holds true:*

$$\|v;H^{l+1}(\Omega)^3\|+\|\nabla p;H^{l-1}(\Omega)^3\|\le C\|f;H^{l-1}(\Omega)^3\|, \tag{7.2}$$

where C is a constant independent of f.

By means of the divergence theorem, it is easy to see that the solutions of Theorem 7.1 produce no flux through the outlet Q_j. With the help of the Poiseuille flow u^{j1} we construct a flux carrier (a divergence-free vector field v_0 with zero trace), which drives a constant prescribed flux H_j through the outlet Q_j in the following way. Suppose v_0 is such a flux carrier; then for every $T>0$

$$\sum_{j=1}^J H_j=\int_{\Gamma_T}\nu^\top\cdot v_0\,\mathrm{d}o=\int_{\Omega_T}\operatorname{div} v_0\,\mathrm{d}x=0.$$

Therefore, $\sum_j H_j=0$ is a necessary (and physically sensible) condition. Let χ_j, $j=1,\dots,J$ be the cut-off functions as in Lemma 1.4, and u^{j1} the Poiseuille flow corresponding to the cylinder $\Pi_j=\omega_j\times\mathbf{R}$. Then $g_0=\operatorname{div}\sum_j(H_j\chi_j v^{j1})=\sum H_j(\nabla\chi_j)v^{j1}\in\overset{\circ}{H}{}^1(\Omega_{R_0})$ for a suitable $R_0>0$. Moreover,

$$\int_{\Omega_{R_0}}g_0\,\mathrm{d}x=\sum_j\int_{\omega_j}\mathrm{e}^\top_{z_j}\cdot v^{j1}(y_j,R_0)\,\mathrm{d}y_j=\sum_{j=1}^J H_j=0.$$

Hence, by using the result of [5] once more, we find $w_0 \in \overset{\circ}{H}{}^2(\Omega_{R_0})^3$ with div $w_0 = g_0$. We remember the notation $\mathcal{U}^1 = (\chi_1 u^{11}, \dots, \chi_J u^{J1})$, where the columns u^{j1} again denote the Poiseuille flow through the outlet Q_j. If we put

$$u_0 = \begin{pmatrix} v_0 \\ p_0 \end{pmatrix} = \mathcal{U}^1 \cdot H + \begin{pmatrix} w_0 \\ 0 \end{pmatrix}, \quad H = (H_1, \dots, H_J),$$

we obtain div $v_0 = 0$ in Ω, $v_0 = 0$ on $\partial\Omega$, and $-\Delta v_0 + \nabla p_0 = f_0 \in L^2(\Omega)^3$ with compact support in Ω_{R_0}. Moreover, from the construction of w_0, the following estimate follows:

$$\|f_0; L^2(\Omega)\| \le C \sum_j |H_j|. \tag{7.3}$$

We find a solution $u = u_0 + \tilde{u}$ to the problem (7.1) such that u has prescribed flux H_j through the outlet Q_j. If we apply Theorem 7.1 to the problem

$$-\Delta \tilde{v} + \nabla \tilde{p} = f - f_0, \quad \text{div } \tilde{v} = 0 \text{ in } \Omega, \quad \tilde{v} = 0 \text{ on } \partial\Omega, \tag{7.4}$$

we can formulate the following result.

COROLLARY 7.2. *Let $\Omega \subset \mathbf{R}^3$ be as in Theorem 7.1. For every $f \in L^2(\Omega)^3$, $H \in \mathbb{C}^J$ with $\sum_j H_j = 0$ we obtain a solution u to the problem (7.1) with flux H_j through Q_j in the form $u = u_0 + \tilde{u}$. Here $\tilde{u}$ is uniquely determined up to a constant in pressure and fulfills the following estimate:*

$$\|\tilde{v}, H^2(\Omega)^3\| + \|\nabla \tilde{p}; L^2(\Omega)^3\| \le C\Big(\|f; L^2(\Omega)^2\| + \sum_{j=1}^{J} |H_j|\Big). \tag{7.5}$$

The last result of this treatment concerns the approximation of the solutions obtained by Corollary 7.2.

THEOREM 7.3. *Let $\Omega \subset \mathbf{R}^3$ be a domain with J cylindrical outlets, $\partial\Omega \in C^3$. For $f \in L^2(\Omega)^3$, $H \in \mathbb{C}^j$ with $\sum_{j=1}^{J} H_j = 0$ let u be the solution of Corollary 7.2 to the problem (7.1) with flux H_j through Q_j. For $R > R_0$, let $u^R \in H^2(\Omega_R)^3 \times H^1(\Omega_R)$ be a solution to the approximation problem*

$$-\Delta v^R + \nabla p^R = f|_{\Omega_R}, \quad \text{div } v^R = 0 \text{ in } \Omega_R, \tag{7.6}$$

$$v^R = 0 \text{ on } \partial\Omega(R), \quad Fu^R = F(\mathcal{U}^1 \cdot H) \text{ on } \Gamma_R. \tag{7.7}$$

Then $\|v - v^R; H^2(\Omega_{R-\varepsilon})^3\| + \|\nabla p - \nabla p^R; L^2(\Omega_{R-\varepsilon})^3\| \to 0$ with $R \to \infty$ for every $\varepsilon > 0$.

Proof. The existence of u^R, uniquely determined up to a constant in pressure, was already proved in Theorem 4.2. Since $\sum_j H_j = 0$, it is not necessary now to introduce the "artificial" flux H_∞, as in Theorem 6.1. We apply Corollary 5.3 to $u - u^R$. By Corollary 7.2, $F(u - u^R) = F\tilde{u}$, the proof of Theorem 7.1 yields $\int_{\Gamma_{R,j}} \nu^\top \cdot F\tilde{u} = 0$ for $j = 1, \dots, J$. Therefore (5.24), the trace estimates, and Lemma 4.6 lead to

$$\begin{aligned} &\|v - v^R; H^2(\Omega_{R-\varepsilon})^3\| + \|\nabla p - \nabla p^R; L^2(\Omega_{R-\varepsilon})^3\| \\ &\le C\|F\tilde{u}; \mathcal{R}_1^1 V(\Gamma_R, \partial\Gamma_R)\| \\ &\le C\|\tilde{v}; \mathcal{D}_1^1 V(\Omega_R \backslash \Omega_{R-\varepsilon})\| \le \|\tilde{v}; H^2(\Omega_R \backslash \Omega_{R-\varepsilon})^3\|. \end{aligned}$$

Here we observe that the mixed boundary condition depends only on the velocity part, not on the pressure. Since $\tilde{v} \in H^2(\Omega_R)^3$, the right-hand side of the last inequality tends to 0, as $R \to \infty$, which proves the theorem. □

Remark 7.4. It is possible to formulate these results also in H^l-spaces for arbitrary $l \in \mathbf{N}$. Then in the formulation of the approximation problem, f has to be replaced by $\chi_R f$, like in Theorem 6.1. However, the corresponding results of Corollary 1.2 need a more detailed and technical discussion of the solution w_0 to the divergence equation; we therefore omit this part.

As already mentioned, without more specified assumptions on the decay properties of f, it is not possible to fix the order of convergence of $u - u^R$. But it is possible to refine the estimates after calculating how, for example, the property $f\mathfrak{z}^\gamma \in L^2(\Omega)^3$, $\gamma > 0$ influences the asymptotic behavior of u. This requires other tools than those explained in section 1.2 and will be done in a forthcoming paper.

Acknowledgment. The author is greatly indebted to Professor Sergueï Nazarov for many helpful discussions and advices.

REFERENCES

[1] R. A. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying boundary conditions* I, Comm. Pure Appl. Math., 12 (1959), pp. 623–727.

[2] R. A. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying boundary conditions* II, Comm. Pure Appl. Math., 17 (1964), pp. 35–92.

[3] C. J. Amick, *Steady solutions of the Navier–Stokes equations in unbounded channels and pipes*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 4 (1977), pp. 473–513.

[4] C. J. Amick, *Properties of steady solutions of the Navier–Stokes equations for certain unbounded channels and pipes*, Nonlinear Anal., 2 (1978), pp. 689–720.

[5] W. Borchers and H. Sohr, *On the equation* rot $v = g$ *and* div $u = f$ *with zero boundary conditions*, Hokkaido Math. J., 19 (1993), pp. 67–87.

[6] L. Cattabriga, *Su un problema al contorno relativo al sistema di equazioni di Stokes*, Sem. Mat. Univ. Padova, 31 (1964), pp. 308–340.

[7] M. Dauge, *Stationary Stokes and Navier–Stokes systems on two- or three-dimensional domains with corners, Part* I: *Linearized equations*, SIAM J. Math. Anal., 20 (1989), pp. 74–97.

[8] P. Deuring, *Finite Element methods for the Stokes system in three–dimensional exterior domains*, Math. Methods Appl. Sci., 20 (1997), pp. 245–269.

[9] M. Van Dyke, *Perturbation Methods in Fluid Mechanics*, Academic Press, New York, 1964.

[10] G. Duvaut and J. L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, 1976.

[11] G. P. Galdi, *An introduction to the mathematical theory of the Navier-Stokes equations*, Vol. I, Springer Tracts in Natural Philosophy, Springer-Verlag, New York, 1994.

[12] D. Givoli, *Non reflecting boundary conditions*, J. Comput. Phys., 94 (1991), pp. 1–29.

[13] V. Girault and A. Sequeira, *A well posed problem for the exterior Stokes equation in two and three dimensions*, Arch. Rational Mech. Anal., 114 (1991), pp. 313–333.

[14] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman Advanced Program, Boston, 1985.

[15] G. H. Guirguis and M. D. Gunzberger, *On the approximation of the exterior Stokes problem in three dimensions*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 445–464.

[16] L. Halpern and M. Schatzmann, *Artificial boundary conditions for incompressible viscous flows*, SIAM J. Math. Anal., 20 (1989), pp. 308–353.

[17] L. Halpern, *Spectral methods in polar coordinates for the Stokes problem. Application to computation in unbounded domains*, Math. Comput., 65 (1996), pp. 507–531.

[18] H. Han, J. Lu, and W. Bao, *A discrete artificial boundary condition for steady incompressible viscous flows in a no-slip channel using a fast iterative method*, J. Comput. Physics, 114 (1994), pp. 201–208.

[19] H. Han and W. Bao, *An artificial boundary condition for the incompressible viscous flow in a no slip channel*, J. Comput. Math., 13 (1995), pp. 51–63.
[20] T. Hagstrom and H. B. Keller, *Exact boundary conditions at an artificial boundary for partial differential equations in cylinders*, SIAM J. Math. Anal., 17 (1986), pp. 322–341.
[21] T. Hagstrom and H. B. Keller, *Asymptotic boundary conditions and numerical methods for nonlinear elliptic problems on unbounded domains*, Math. Comp., 48 (1987), pp. 449–470
[22] J. G. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations*, Intern. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.
[23] L. V. Kapitanski and K. Pileckas, *On spaces of solenoidal vector fields and boundary value problems for the Navier-Stokes equations in domains with noncompact boundaries*, Trudy mat. Inst. Steklov, 159 (1983), pp. 5–36; Proc. Math. Inst. Steklov, 159, (1984), pp. 3–34 (in English).
[24] L. V. Kapitanski and K. Pileckas, *Certain problems of vector analysis*, Zapiski Nauchn. Sem. LOMI, 138 (1984), pp. 65–85 (in Russian); J. Sov. Math., 32 (1986), pp. 469–483 (in English).
[25] R. B. Kellogg and J. E. Osburn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.
[26] O. A. Ladyženskaja, *The Mathematical Theory of Viscous Incompressible Flow*, Gordon & Breach, New York, 1966.
[27] O. A. Ladyženskaja and V. A. Solonnikov, *Determination of the solutions of boundary value problems for steady Stokes and Navier-Stokes equations having an unbounded Dirichlet integral*, Zapiski Nauchn. Sem. LOMI, 96 (1980), pp. 117–160; J. Sov. Math., 21 (1983), pp. 728–761 (in English).
[28] J. L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. I, Springer Verlag, Berlin, 1972
[29] V. G. Maz'ja and B. A. Plamenevskiĭ, *Estimates in L_p and Hölder classes and the Miranda-Agmon maximum principle for solutions of elliptic boundary value problems in domains with singular points on the boundary*, Math. Nachr., 81 (1978), pp. 25–82 (in Russian); Amer. Math. Soc. Transl. Ser. 2, 123 (1984), pp. 1–56 (in English).
[30] V. G. Maz'ja, S. A. Nazarov, and B. A. Plamenevskiĭ, *On asymptotics of solutions to elliptic boundary value problems under irregular perturbations of domains*, Problemy Matem. Analisa Leningrad University, 8 (1981), pp. 72–153 (in Russian).
[31] V. G. Maz'ja, S. A. Nazarov, and B. A. Plamenevskiĭ, *Asymptotische Theorie elliptischer Randwertaufgaben in singulär gestörten Gebieten*, Akademie Verlag, Berlin, 1991
[32] V. G. Maz'ja, B. A. Plamenevskiĭ, and L. Stupelis, *Three-dimensional problem with a free boundary*, Diff. equations and their applications, Inst. of Math. and Cybern. Acad. Sci. Lit. SSR, Vilnius, (1979), pp. 9–153 (in Russian).
[33] S. A. Nazarov and B. A. Plamenevskiĭ, *Elliptic Problems in Domains with Piecewise Smooth Boundaries*, De Gruyter Verlag, Berlin, 1994.
[34] S. A. Nazarov, *On the asymptotics with respect to a parameter of the solution of a boundary value problem with periodic coefficients in a cylinder*, Differ. Uravn. Primen., 30 (1981), pp. 27–45 (in Russian).
[35] S. A. Nazarov and M. Specovius-Neugebauer, *Approximation of exterior boundary value problems for the Stokes system*, Asymptot. Anal., 14 (1997), pp. 223–255.
[36] S. A. Nazarov and K. Pileckas, *Asymptotic conditions at infinity for the Stokes and Navier-Stokes problems in domains with cylindrical outlets to infinity*, Acta Appl. Math., to appear.
[37] A. Passerini and G. Thäter, *The Stokes system in domains with outlets of bounded cross-section containing a cylinder*, Z. Anal. Anwendungen, 17 (1998), pp. 615–639.
[38] A. Pazy, *Asymptotic expansion of solutions of ordinary differential equations in Hilbert spaces*, Arch. Rational Mech. Anal., 24 (1967), pp. 193–218.
[39] K. Pileckas *On spaces of solenoidal vectors*, Trudy Mat. Inst. Steklov, 159 (1983), pp. 137–149 (in Russian); Proc. Mat. Inst. Steklov, 159 (1984), pp. 141–154 (in English).
[40] K. Pileckas *Weighted L^q–theory, uniform estimates and asymptotics for steady Stokes and Navier–Stokes equations in domains with noncompact boundaries*, Habilitation thesis, Fachbereich 17 Mathematik-Informatik, University of Paderborn, Paderborn 1994.
[41] Y. A. Roĭtberg and Z. G. Sheftel', *A theorem of homeomorphisms for elliptic systems and its applications*, Math. USSS–Sb., 7 (1969), pp. 439–465.
[42] Y. A. Roĭtberg, *Boundary values of generalized solutions of systems that are elliptic in the sense of Douglis and Nirenberg*, Siberian Math. J., 18 (1977), pp. 600–610.
[43] Y. A. Roĭtberg, *Elliptic Boundary Value Problems in the Spaces of Distributions*, Kluwer

Academic Publishers, Dordrecht, 1996.

[44] V. S. Ryabenki and S. V. Tsynkov, *Artificial boundary conditions for the numerical solution of external viscous flow problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1355–1389.

[45] A. M. Sändig and M. Orlt, *Regularity of viscous Navier–Stokes flows in nonsmooth domains*, in Boundary Value Problems and Integral Equations in Nonsmooth Domains, M. Costabel, M. Dauge, and S. Nicaise, eds., Lecture Notes in Pure and Appl. Math. 167, Dekker, New York, 1995.

[46] A. Sequeira, *The coupling of boundary integral and finite element methods for the bidimensional steady Stokes problem*, Math. Methods Appl. Sci., 5 (1983), pp. 356–376.

[47] V. A. Solonnikov, *On the boundary value problems for systems elliptic in the sense of A. Douglis, L. Nirenberg*, Amer. Math. Soc. Transl., 56 (1966), pp. 192–232.

[48] V. A. Solonnikov and V. E. Scadilov, *On a boundary value problem for a steady system of Navier–Stokes equations*, Proc. Mat. Inst. Steklov, 125 (1973), pp. 186–199.

[49] V. A. Solonnikov, *On the Stokes equations in domains with non-smooth boundaries and on viscous incompressible flow with a free surface*, College de France Seminar 3, Res. Notes Math., 70 (1982), pp. 340–423.

[50] V. I Smirnov, *Lehrgang der höheren Mathematik*, Teil IV, 6th ed., VEB Verlag der Wissenschaften, Berlin, 1973.

[51] I. L. Sofronov, *Non-reflecting inflow and outflow in a wind tunnel for transonic time-accurate simulation*, J. Math. Anal. Appl., 221 (1998), pp. 92–115.

[52] M. Specovius–Neugebauer, *Approximation of Stokes problems in unbounded domains*, Habilitation Thesis, Fachbereich 17 Mathematik-Informatik, University of Paderborn, Paderborn, 1997.

[53] S. V Tsynkov, *Artificial boundary conditions based on the difference potential method*, NASA Technical Memorandum 110265, Langley Research Center, Hampton, Virginia, 1996.

[54] M. L. Williams, *Stress singularities resulting from various boundary conditions in angular corners of plates in extension*, J. Appl. Mech., 19 (1952), pp. 526–528.

SIAM J. MATH. ANAL.
Vol. 30, No. 3, pp. 678–697

ARBITRARILY SMOOTH ORTHOGONAL NONSEPARABLE WAVELETS IN $\mathbb{R}^2$*

EUGENE BELOGAY† AND YANG WANG†

Abstract. For each $r \in \mathbb{N}$, we construct a family of bivariate orthogonal wavelets with compact support that are nonseparable and have vanishing moments of order r or less. The starting point of the construction is a scaling function that satisfies a dilation equation with special coefficients and a special dilation matrix M: the coefficients are aligned along two adjacent rows, and $|\det(M)| = 2$. We prove that if $M^2 = \pm 2I$, e. g., $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ or $M = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, then the smoothness of the wavelets improves asymptotically by $1 - \frac{1}{2}\log_2 3 \approx 0.2075$ when r is incremented by 1. Hence they can be made arbitrarily smooth by choosing r large enough.

Key words. nonseparable wavelets, smooth orthogonal scaling function, regularity

AMS subject classifications. Primary, 42C15; Secondary, 26B35, 41A25, 41A63, 65D20

PII. S0036141097327732

1. Introduction.

Since the introduction by Daubechies [7] of compactly supported orthogonal wavelet bases in $\mathbb{R}^1$ with arbitrarily high smoothness, various new wavelet bases (often with specially tailored properties) have been constructed and applied successfully in image processing, numerical computation, statistics, etc. Many of these applications, such as image compression, employ wavelet bases in $\mathbb{R}^2$. Virtually all of these bases are *separable*; that is, the bivariate basis functions are simply tensor products of univariate basis functions. A separable wavelet basis is easy to construct and simple to study, for it inherits features of the corresponding wavelet basis in $\mathbb{R}^1$, such as smoothness and support size. Separable wavelet transforms are easy to implement.

Nevertheless, separable bases have a number of drawbacks. Because they are so special, they have very little design freedom. Furthermore, separability imposes an unnecessary product structure on the plane, which is artificial for natural images. For example, the zero set of a separable scaling function contains horizontal and vertical lines. This "preferred directions" effect can create unpleasant artifacts that become obvious at high image compression ratios. Nonseparable wavelet bases offer the hope of a more isotropic analysis [6, 12, 15].

Despite the success in constructing univariate orthogonal and multivariate biorthogonal wavelet bases with arbitrarily high smoothness, a general theory of smooth multivariate orthogonal nonseparable wavelets is not currently available, and only a few such constructions have been published. Gröchenig and Madych [9] constructed several nonseparable Haar-type scaling functions in $\mathbb{R}^n$, which are discontinuous indicator functions of (often fractal-like) compact sets. Cohen and Daubechies [6] used the univariate construction [7] to produce nonseparable scaling functions with higher accuracy, which are not continuous, as proved by Villemoes [16]. Continuous nonseparable scaling functions were constructed by Kovačević and Vetterli [12], and recently by He and Lai [10], but none of these functions is differentiable.

*Received by the editors September 22, 1997; accepted for publication (in revised form) May 26, 1998; published electronically April 9, 1999.
http://www.siam.org/journals/sima/30-3/32773.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (ebelogay@math.gatech.edu, wang@math.gatech.edu).

In this paper we present a new family of arbitrarily smooth, orthogonal, nonseparable wavelet bases in $\mathbb{R}^2$. Our construction follows the standard multiresolution analysis (MRA) approach [8, 13, 14]: it focuses on a scaling function that solves a dilation (or refinement) equation with special coefficients and a special dilation matrix M with $|\det(M)| = 2$. Such dilation matrices make a popular "laboratory case," partly because the MRA involves only one wavelet [4, 6, 12, 15]. The wavelet is easy to construct from the scaling function and has the same smoothness, so we deal mainly with the scaling function.

The coefficients in the dilation equation, called scaling coefficients, determine the properties of the scaling function. To construct a scaling function usually means to find its scaling coefficients. We characterize completely the set of two-row scaling coefficients that produce nonseparable orthogonal scaling functions with arbitrarily high accuracy. We show that our construction can produce scaling functions of any desired smoothness for the special dilation matrix $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$.

The paper is organized as follows. In section 2 we introduce some basic notions and assumptions and state our two main results: the first describes the scaling coefficients; the second determines the smoothness of the scaling function. In section 3 we formulate and solve the equations that the scaling coefficients must satisfy in order for the scaling function to be orthogonal and accurate. In section 4 we prove the smoothness result. In section 5 we plot some of the new scaling functions and explain how our results can be formulated for dilation matrices other than $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$.

Note. After submitting this paper, the authors learned about related independent results by Ayache [1], who recently constructed arbitrarily smooth nonseparable orthogonal wavelets by perturbing the separable wavelets for dilation $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$.

2. Preliminaries and main results. Let M be an expanding 2×2 matrix with integer entries such that $|\det(M)| = 2$. The key ingredients to an MRA with such a *dilation matrix* M are two functions: a *scaling function* ϕ and a *wavelet* ψ. The scaling function $\phi : \mathbb{R}^2 \to \mathbb{R}$ satisfies a *dilation equation* (*refinement equation*) of the form

$$\phi(\mathbf{x}) = 2 \sum_{\mathbf{n} \in \mathbb{Z}^2} c_{\mathbf{n}} \phi(M\mathbf{x} - \mathbf{n}), \qquad \text{where} \quad \sum_{\mathbf{n} \in \mathbb{Z}^2} c_{\mathbf{n}} = 1. \tag{2.1}$$

The numbers $c_{\mathbf{n}}$ are the *scaling coefficients* (*low-pass filter coefficients*) of $\phi(\mathbf{x})$. We assume that they are real and that $c_{\mathbf{n}} \neq 0$ for only finitely many $\mathbf{n} \in \mathbb{Z}^2$ (ensuring that $\phi(\mathbf{x})$ has compact support). A convenient way to work with the scaling coefficients as a whole is to consider the *coefficient mask* (z-*transform*)

$$C(z, w) := \sum_{(m,n) \in \mathbb{Z}^2} c_{(m,n)} z^m w^n, \quad \text{where } (z, w) \in \mathbb{C}^2.$$

Note that $C(1, 1) = 1$. The *symbol* (*frequency response*) of (2.1) is

$$m(\omega_1, \omega_2) := C(e^{i\omega_1}, e^{i\omega_2}).$$

Denote M^{T} by W. The mask is said to satisfy *Cohen's criterion* [16, p. 181] if there exists a compact fundamental domain Ω of the lattice $2\pi\mathbb{Z}^2$ with the property

$$m(W^{-j}\omega) \neq 0 \quad \text{for all } j \geq 1 \text{ and for all } \omega \in \Omega. \tag{2.2}$$

An important task in wavelet theory is to relate the properties of the scaling function $\phi(\mathbf{x})$ to the properties of the coefficient mask $C(z, w)$. Our goal is to find

coefficients in (2.1) that produce a scaling function with two important properties: orthogonality and accuracy.

A scaling function $\phi(\mathbf{x}) \in L^2(\mathbb{R}^2)$ is called *orthogonal* if the set of its lattice translates $\{\phi(\mathbf{x}-\mathbf{k}) : \mathbf{k} \in \mathbb{Z}^2\}$ is orthogonal. The following *orthogonal coefficients condition* is necessary for $\phi(\mathbf{x})$ to be orthogonal:

$$2\sum_{\mathbf{n}\in\mathbb{Z}^2} c_{\mathbf{n}} c_{\mathbf{n}+M\mathbf{k}} = \delta_{\mathbf{0},\,\mathbf{k}} \qquad \text{for all } \mathbf{k} \in \mathbb{Z}^2, \tag{2.3}$$

where $\delta_{\mathbf{m},\,\mathbf{k}}$ is the Kronecker symbol. The condition (2.3) becomes sufficient if, in addition, the scaling coefficients satisfy Cohen's criterion (2.2). The coefficient mask $C(z,w)$ is called *orthogonal* if (2.3) holds.

A scaling function $\phi(\mathbf{x})$ is said to have *accuracy* (*regularity*) r if the space of the infinite linear combinations of $\{\phi(\mathbf{x}-\mathbf{k}) : \mathbf{k} \in \mathbb{Z}^2\}$ contains all polynomials of degree $r-1$ or less. In this case the coefficient mask $C(z,w)$ is said to have accuracy r as well. (How the accuracy of ϕ relates to $C(z,w)$ is explained in the next section.) Accuracy is a desirable property in many applications; for example, in image processing it implies that the polynomial components of the filtered signal will not "leak" into the high-pass band, which improves compression.

In this paper we measure the smoothness of a real bivariate function ϕ by its Hölder exponent (there are alternative regularity measures, such as the Sobolev exponent). Let $s = m + \gamma$, where $0 \le \gamma < 1$ and m is a nonnegative integer. Then we say that $\phi(\mathbf{x}) \in C^s$ if for each partial derivative $D^\alpha\phi(\mathbf{x})$, where $\alpha = (\alpha_1, \alpha_2)$ and $\alpha_1 + \alpha_2 \le m$, there is a constant $c > 0$ such that $|D^\alpha\phi(\mathbf{x}) - D^\alpha\phi(\mathbf{y})| \le c|\mathbf{x}-\mathbf{y}|^\gamma$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. If the Fourier transform $\widehat{\phi}(\omega)$ of $\phi(\mathbf{x})$ satisfies

$$|\widehat{\phi}(\omega)| \le c \cdot (1+|\omega_1|)^{-s-1-\epsilon}(1+|\omega_2|)^{-s-1-\epsilon}$$

for some constants $c > 0$ and $\epsilon > 0$, then $\phi(\mathbf{x}) \in C^s$. It is usually harder to estimate the smoothness of a scaling function than its accuracy.

The *autocorrelation* (*product filter*) of a real polynomial $F(z_1, z_2)$ is defined to be

$$\mathcal{P}_F(z_1, z_2) := F(z_1, z_2)F(z_1^{-1}, z_2^{-1}).$$

The polynomial F is called a *spectral factor* of $\mathcal{P}_F$. Note that $\mathcal{P}_F(e^{i\omega_1}, e^{i\omega_2}) = |F(e^{i\omega_1}, e^{i\omega_2})|^2$ and is therefore always real and nonnegative. Conversely, the Fèjer–Riesz theorem [7] ensures that every real-valued nonnegative univariate trigonometric polynomial $P(e^{i\omega_1})$ can be factored (nonuniquely) as $|F(e^{i\omega_1})|^2$. In most cases the spectral factor F can only be computed numerically. (Strang and Nguyen [15, p. 157] review several spectral factorization methods.)

Spectral factorization is a key technique in univariate orthogonal wavelet theory. Daubechies [7] constructed orthogonal univariate wavelets of arbitrarily high accuracy by deriving an analytic formula for the autocorrelation of the coefficient mask; the scaling coefficients themselves must be computed numerically. Theorem 2.1 is a similar result in a special bivariate setting.

The main difficulty in constructing bivariate nonseparable smooth wavelets is that some key univariate techniques, such as polynomial factorization, do not generalize to the bivariate case. The Fèjer–Riesz theorem is one of them: e.g., $2+\cos(\omega_1)+\cos(\omega_2)$ is real and nonnegative; yet it cannot be factored as $|F(e^{i\omega_1}, e^{i\omega_2})|^2$. One approach to nonseparable orthogonal wavelets is to solve the accuracy and orthogonality conditions for a specific arrangement of unknown scaling coefficients. Unfortunately, the resulting

system of linear and quadratic equations can only be solved (even numerically) when the number of coefficients is rather small. (Kovačević and Vetterli [12] placed the 8 nonzero coefficients in the $\bullet\!\!:\!\!:\!\!\bullet$ pattern; He and Lai [10] used a 4×4 mask.) Another approach to nonseparable wavelets, which we adopt here, is to study special cases of (2.1) and derive the scaling coefficients explicitly.

Our construction employs a special coefficient mask for a particular dilation matrix (generalizations are discussed in section 5). First, we fix the dilation matrix

$$M = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix}.$$

The property $M^2 = 2I$ will be useful in the proof of Theorem 2.2. Next, we restrict the placement of the nonzero scaling coefficients in (2.1) to two adjacent rows in the first quadrant. That is, $\operatorname{supp} c := \{\mathbf{n} \in \mathbb{Z}^2 : c_{\mathbf{n}} \neq 0\} \subseteq \{0, \dots, N\} \times \{0, 1\}$ for some $N \in \mathbb{N}$. As a result, the coefficient mask has the form

$$C(z, w) = A(z) + wB(z), \tag{2.4}$$

where $A(z)$ and $B(z)$ are polynomials of one complex variable. The theorems below do not generalize easily to masks with more than two rows.

Observe that $A(1)+B(1) = C(1,1) = 1$, so it is impossible that $A(1) = B(1) = 0$. Therefore, we assume that $A(1) \neq 0$, for we can always achieve it by a suitable change of variables in (2.1), as explained at the end of section 3. Furthermore, we assume that $B(0) \neq 0$. Cohen and Daubechies [6] noted that if $B(z) \equiv 0$ (that is, if the scaling coefficients are aligned along a single horizontal line), then the scaling function is separable. A simple but important example of such a one-row mask is the *Haar coefficient mask*

$$H(z) := \frac{1}{2}(1 + z). \tag{2.5}$$

The corresponding separable scaling function is the indicator of the unit square.

Our first theorem gives the necessary conditions for a two-row coefficient mask to produce a nonseparable orthogonal scaling function of arbitrarily high accuracy. The proof is in section 3.

THEOREM 2.1. *Let $\phi(\mathbf{x})$ satisfy the dilation equation (2.1) with dilation $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ and coefficient mask $C(z, w) = A(z) + wB(z)$, where $A(1) \neq 0$ and $B(0) \neq 0$. Let $r \in \mathbb{N}$ and let ν be an odd integer with $\nu \geq \deg A$. If the scaling function $\phi(\mathbf{x})$ is orthogonal and has accuracy $r+1$, then the polynomials $A(z)$ and $B(z)$ have the form*

$$z^\nu A(z^{-1}) = H^r(z)L(z)S(z^2), \tag{2.6}$$

$$B(z) = H^r(z)L(-z)Q(z^2)H^{2r}(-z), \tag{2.7}$$

where $L(z)$, $S(z)$, and $Q(z)$ are any polynomials that satisfy

$$\mathcal{P}_S(z^2) = 1 - \left(\frac{1-z^2}{4}\right)^r \left(\frac{1-z^{-2}}{4}\right)^r \mathcal{P}_Q(z^2), \tag{2.8}$$

$$\mathcal{P}_L(z) = \sum_{j=0}^{r-1} \binom{r+j-1}{j} \left(\frac{1-u}{2}\right)^j + (1-u)^r u\, R(u^2), \quad u := \frac{1}{2}(z + z^{-1}), \tag{2.9}$$

$$S(1) = L(1) = 1, Q(1) = (-1)^r L(-1), \quad \text{and} \quad L(0)Q(0) \neq 0 \tag{2.10}$$

and $R(z)$ is an arbitrary polynomial.

Remark. The converse of Theorem 2.1 holds if, in addition to (2.6)–(2.10), the mask $C(z, w)$ satisfies Cohen's criterion (2.2).

Theorem 2.1 suggests the following procedure for obtaining the coefficients of $C(z, w)$:

(i) Choose polynomials Q and R so that the right-hand sides of (2.8)–(2.9) are nonnegative along the unit circle $|z| = 1$ (thus ensuring that the next step can be performed).

(ii) Using some spectral factorization method, make a list of all polynomials S and L that satisfy (2.8)–(2.10). Choose a specific pair. (Observe that if $R = 0$, then the polynomial L is one of the univariate orthogonal coefficients masks with accuracy r constructed by Daubechies [7].)

(iii) Substitute L, S, and Q in (2.6)–(2.7) and expand. Choose an odd $\nu \geq \deg A$.

Note that the choices in step (ii) are not unique. Unlike the univariate case, there is no obvious way to single out a "minimal phase" coefficient mask $C(z, w)$.

The minimal degree of $A(z)$ and $B(z)$ in Theorem 2.1 is $4r - 1$ and is achieved if

$$Q(z) = \text{const} = (-1)^r L(-1), \quad R(z) = \text{const} = 0, \quad \text{and} \quad \nu = 4r - 1. \tag{2.11}$$

There are $2^{1+2\lfloor r/2 \rfloor}$ coefficient masks $C(z, w)$ that satisfy the conditions (2.11) and $A(0) \neq 0$ in addition to the conditions in Theorem 2.1. Denote the set of those masks by $\mathcal{C}_{r+1}$. Every $C(z, w)$ in $\mathcal{C}_{r+1}$ produces a scaling function $\phi(\mathbf{x})$ with accuracy $r + 1$. Denote the set of those scaling functions by Φ_{r+1}.

Our second theorem states that all functions in Φ_{r+1} are compactly supported orthogonal nonseparable scaling functions that can be made arbitrarily smooth by choosing the accuracy $r + 1$ large enough. The proof is in section 4.

THEOREM 2.2. *Let* $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ *and let* $\phi \in \Phi_{r+1}$. *Then*

(i) ϕ *is an orthogonal scaling function with accuracy* $r + 1$;

(ii) $\operatorname{supp} \phi \subseteq [0, 4r + 1] \times [0, 4r]$;

(iii) ϕ *is nonseparable;*

(iv) *if* $r \geq 5$, *then* $\phi \in C^{r - \mu_r - 2}$, *where*

$$\mu_r := \frac{1}{2} \log_2 \left(\sum_{j=0}^{r-1} \binom{r + j - 1}{j} \left(\frac{3}{4}\right)^j \right).$$

Furthermore,

$$r - \mu_r > \left(1 - \frac{1}{2} \log_2 3\right) r + \frac{1}{2} \log_2 3 \tag{2.12}$$

and

$$\lim_{r \to \infty} \frac{r - \mu_r - 2}{r} = 1 - \frac{1}{2} \log_2 3 \approx 0.2075. \tag{2.13}$$

Remark. As in the univariate case, the statements (2.12)–(2.13) guarantee the existence of orthogonal wavelets of any desired smoothness. In particular, $\Phi_{5+1} \subset C^0$, $\Phi_{9+1} \subset C^1$, and $\Phi_{13+1} \subset C^2$. The smoothness estimate $\phi \in C^{r - \mu_r - 2}$ applies to all ϕ in Φ_{r+1} uniformly and is not sharp. Lemma 5.1 provides sharper smoothness estimates for individual functions in Φ_{r+1}, involving numerical computations.

If $\phi(\mathbf{x})$ is an orthogonal scaling function, then its associated *wavelet* $\psi(\mathbf{x})$ has the form

$$\psi(\mathbf{x}) = 2 \sum_{\mathbf{n} \in \mathbb{Z}^2} d_{\mathbf{n}} \phi(M\mathbf{x} - \mathbf{n}).$$

FIG. 2.1. *The quincunx and the column sublattice.*

Obviously, the wavelet ϕ has the same smoothness as the scaling function ϕ. It is known that if the *wavelet coefficients* (*high-pass filter coefficients*) $d_{\mathbf{n}}$ are given by

$$d_{\mathbf{n}} = (-1)^{n_1} c_{\mathbf{e}-\mathbf{n}}, \quad \text{where } \mathbf{n} = (n_1, n_2) \text{ and } \mathbf{e} = (1, 0),$$

then the wavelet $\psi(\mathbf{x})$ is orthogonal to $\{\phi(\mathbf{x} - \mathbf{k}) : \mathbf{k} \in \mathbb{Z}^2\}$, and its lattice translates and dilates form an orthogonal basis of $L^2(\mathbb{R}^2)$ [13, 14].

Although Theorems 2.1 and 2.2 are formulated for $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$, they can be generalized [2] to other expanding matrices with integer entries and $|\det(M)| = 2$. Each such matrix transforms the lattice $\mathbb{Z}^2$ into three types of sublattices: the *quincunx*, the *column* sublattice (see Figure 2.1), and the *row* sublattice (the row sublattice is merely a transpose of the column sublattice).

The accuracy and the orthogonality conditions depend only on the sublattice type and not on the dilation matrix, as we shall see in section 3. Therefore, Theorem 2.1 still holds if $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ is replaced by any other expanding integral matrix M that generates the column sublattice $2\mathbb{Z} \times \mathbb{Z}$, such as $\left(\begin{smallmatrix} 0 & -2 \\ 1 & 0 \end{smallmatrix}\right)$. If $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ is replaced by the quincunx matrix $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$, or any other matrix that generates the quincunx lattice, such as $\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, then $A(z) + wB(z)$ must be replaced by $A(z) + wzB(z)$, as we shall see at the end of section 5.

In contrast, the smoothness of the scaling function does depend on the dilation matrix. Theorem 2.2 can be extended to any dilation M with the properties $M^2 = \pm 2I$, such as $\left(\begin{smallmatrix} 0 & -2 \\ 1 & 0 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$. Although the quincunx matrix $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ has often been used in nonseparable constructions, the authors know of no scaling functions that are both differentiable and orthogonal with respect to the quincunx matrix; it is quite possible that $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ admits no such scaling functions at all.

3. Accuracy and orthogonality conditions. In this section we prove Theorem 2.1. First, we study how the accuracy and the orthogonality of the scaling function restrict the scaling coefficients. Then, we derive the formulas (2.6)–(2.10). Recall that $H(z) = \frac{1}{2}(1 + z)$.

Accuracy. The accuracy of the scaling function ϕ has many equivalent formulations, such as the vanishing moments condition for the wavelet ($\int \mathbf{x}^\alpha \psi(\mathbf{x})\, d\mathbf{x} = 0$, $|\alpha| < r$) or the "sum rules" on the coefficients $c_{\mathbf{n}}$ [4]. Here, we prefer to work with the sum rules. Applied to our dilation matrix $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$, a general result by Cabrelli, Heil, and Molter [4] states that an orthogonal scaling function ϕ has accuracy r if and only if its coefficient mask $C(z, w)$ satisfies the following *accuracy condition*:

$$\frac{\partial^{p+q}}{\partial z^p \partial w^q} C(-1, 1) = 0 \quad \text{for all } p, q \geq 0 \text{ with } p + q < r. \tag{3.1}$$

Observe that the analogous condition for a univariate polynomial $C(z)$ implies the factorization $C(z) = (1+z)^r Q(z)$ for some $Q(z)$. Unfortunately, bivariate polynomials $C(z, w)$ cannot be described in such a simple way. (This fact is one of the main

obstacles in studying nonseparable bivariate wavelets.) Nevertheless, the accuracy condition (3.1) takes on a simple form for our special two-row mask $C(z,w)$.

LEMMA 3.1. *Let* $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$. *The dilation equation* (2.1) *with coefficient mask* $C(z,w) = A(z) + wB(z)$ *has accuracy* $r+1$ *if and only if*

$$A(z) = H^r(z)A_0(z), \quad B(z) = H^r(z)B_0(z), \tag{3.2}$$

$$A_0(-1) + B_0(-1) = 0. \tag{3.3}$$

Proof. It is easy to check that (3.2)–(3.3) imply (3.1). We now prove the converse. By taking $q = 0$ in (3.1), we obtain $A^{(p)}(-1) + B^{(p)}(-1) = 0$ for all $0 \le p \le r$. By taking $q = 1$, we obtain $B^{(p)}(-1) = 0$ for all $0 \le p \le r-1$. These conditions imply $A^{(p)}(-1) = 0$ for all $0 \le p \le r-1$. Hence $A(z) = H^r(z)A_0(z)$ and $B(z) = H^r(z)B_0(z)$. Finally, $A^{(r)}(-1) + B^{(r)}(-1) = 0$ yields $A_0(-1) + B_0(-1) = 0$. ∎

Note that, unlike the univariate case, here accuracy $r+1$ corresponds to r Haar factors plus one extra condition (3.3).

Orthogonality. If $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$, the orthogonality condition (2.3) is equivalent to

$$\mathcal{P}_C(z,w) + \mathcal{P}_C(-z,w) = 1.$$

For the special two-row coefficient mask (2.4), this condition splits further into the following two conditions imposed on the polynomials A and B:

$$\mathcal{P}_A(z) + \mathcal{P}_A(-z) + \mathcal{P}_B(z) + \mathcal{P}_B(-z) = 1, \tag{3.4}$$

$$A(z^{-1})B(z) + A(-z^{-1})B(-z) = 0. \tag{3.5}$$

Remark. The second condition means that $A(z^{-1})B(z)$ contains no even powers of z. Therefore, $\deg A$ must be odd if $B(0) \neq 0$.

The main task in the remainder of this section is to solve (3.2)–(3.5) for the unknown polynomials A and B. Observe that (3.2)–(3.3) are linear and that (3.4)–(3.5) are quadratic.

First, we investigate (3.5).

LEMMA 3.2. *For every polynomial* $p(z) = z^{2n}p_1(z)$ *with* $p_1(0) \neq 0$ *and* $n \ge 0$, *there exist polynomials* q *and* ℓ *such that* $p(z) = q(z^2)\ell(z)$ *and* $\gcd(\ell(z), \ell(-z)) = 1$.

Proof. Consider $t(z) := \gcd(p(z), p(-z))$. Clearly, $t(z) = t(-z) = q(z^2)$ for some polynomial $q(z)$. Let $\ell(z) := p(z)/q(z^2)$, and observe that $\gcd(\ell(z), \ell(-z)) = 1$. ∎

LEMMA 3.3. *The following are equivalent:*

(i) *The polynomials* A *and* B *satisfy condition* (3.5) *and* $B(0) \neq 0$.

(ii) *There exist an odd integer* $\nu \ge \deg A$ *and real polynomials* s, q, *and* ℓ *such that* $z^\nu A(z^{-1}) = s(z^2)\ell(z)$, $B(z) = q(z^2)\ell(-z)$, *and* $\gcd(\ell(z), \ell(-z)) = 1$.

Proof. The implication (ii) ⇒ (i) follows from

$$z^\nu s(z^2)\ell(z)q(z^2)\ell(-z) + (-z)^\nu s(z^2)\ell(-z)q(z^2)\ell(z) = 0.$$

Assume that (i) holds, and fix an odd $\nu \ge \deg A$. Lemma 3.2 allows us to write $z^\nu A(z^{-1}) = s(z^2)\ell(z)$ and $B(z) = q(z^2)m(z)$ for some polynomials s, q, ℓ, and m. Substituting $A(z^{-1})$ and $B(z)$ in (3.5) yields

$$m(z)\ell(z) = m(-z)\ell(-z).$$

Since $\gcd(m(z), m(-z)) = \gcd(\ell(z), \ell(-z)) = 1$, it follows that $m(z) = \ell(-z)$. ∎

Next, we take into account the accuracy of $C(z,w)$.

LEMMA 3.4. *Let A and B be polynomials with $A(1) \neq 0$ and $B(0) \neq 0$. Then the following are equivalent:*

(i) *The two-row coefficient mask $C(z,w) = A(z) + wB(z)$ satisfies* (3.5) *and has accuracy $r+1$.*

(ii) *There exists an odd integer $\nu \geq \deg A$ and real polynomials S, Q, and L with $L(0)Q(0) \neq 0$, $S(1) = L(1) = 1$, and $Q(1) = (-1)^r L(-1)$ such that*

$$z^\nu A(z^{-1}) = H^r(z)L(z)S(z^2),$$
$$B(z) = H^r(z)L(-z)Q(z^2)H^{2r}(-z).$$

Proof. The implication (ii)⇒(i) follows from Lemmas 3.1 and 3.3 by letting $s(z^2) = S(z^2)$, $q(z^2) = Q(z^2)\big(\frac{1-z^2}{2}\big)^r$, and $\ell(z) = H^r(z)L(z)$. We now prove that (i)⇒(ii). By Lemma 3.1, $H^r(z)$ divides both $A(z)$ and $B(z)$. Therefore, $z = -1$ is a root of multiplicity $2r$ of $P(z) := z^\nu A(z^{-1})B(z)$. On the other hand, $P(z) = P(-z)$ by (3.5); hence $z = 1$ is another root of multiplicity $2r$ of $P(z)$. Since $A(1) \neq 0$ by assumption, $H^{2r}(-z)$ must divide $B(z)$. Define

$$A_1(z) := A(z)z^r/H^r(z), \quad \text{and} \quad B_1(z) := B(z)/\big(H^r(z)H^{2r}(-z)\big).$$

Noting that $H(z^{-1}) = z^{-1}H(z)$, we rewrite

$$A(z^{-1})B(z) = H(z)^{2r}H(-z)^{2r}A_1(z^{-1})B_1(z)$$

and substitute in (3.5). Factoring out $H(z)^{2r}H(-z)^{2r}$ yields the equation

$$A_1(z^{-1})B_1(z) + A_1(-z^{-1})B_1(-z) = 0.$$

Therefore, Lemma 3.2 applies to A_1 and B_1, and we obtain the desired factorization in (ii). Since $B(1) = 0$, and therefore $A(1) = 1$, we can normalize L and S so that $L(1) = S(1) = 1$. The only restriction on Q, that $Q(1) = (-1)^r L(-1)$, then follows directly from the extra accuracy condition (3.3). ∎

Finally, to prove Theorem 2.1 we need to solve the remaining equation, (3.4).

Proof of Theorem 2.1. We need only show that the polynomials S, Q and L in Lemma 3.4 have the autocorrelations given by (2.8) and (2.9). By Lemma 3.4,

$$\mathcal{P}_A(z) = \mathcal{P}_H^r(z)\mathcal{P}_S(z^2)\mathcal{P}_L(z), \quad \text{and} \quad \mathcal{P}_B(z) = \mathcal{P}_H^r(z)\mathcal{P}_H^{2r}(-z)\mathcal{P}_L(-z)\mathcal{P}_Q(z^2).$$

Hence $\mathcal{P}_A(z) + \mathcal{P}_B(-z) = \mathcal{P}_H^r(z)\mathcal{P}_L(z)U(z^2)$, where

$$U(z^2) := \mathcal{P}_S(z^2) + \mathcal{P}_H^r(-z)\mathcal{P}_H^r(z)\mathcal{P}_Q(z^2).$$

Therefore, the orthogonality condition (3.4) can be factored as follows:

$$\big(\mathcal{P}_H^r(z)\mathcal{P}_L(z) + \mathcal{P}_H^r(-z)\mathcal{P}_L(-z)\big)U(z^2) = 1.$$

It follows that $U(z^2)$ must be a constant. In particular, $U(z^2) = U(1) = \mathcal{P}_S(1) = 1$, which immediately yields (2.8). Finally, the equation

$$\mathcal{P}_H^r(z)\mathcal{P}_L(z) + \mathcal{P}_H^r(-z)\mathcal{P}_L(-z) = 1 \tag{3.6}$$

characterizes all univariate orthogonal masks with accuracy r and was solved by Daubechies [7]. The solution to (3.6) is given by (2.9). ∎

Remark. Although the results in this section are formulated for $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$, the accuracy and orthogonality conditions (2.3) and (3.1) depend only on the sublattice $M\mathbb{Z}^2$ (in our case, the column sublattice), not on the particular dilation matrix. Therefore, Theorem 2.1 and all results in this section apply to any other integer expanding matrix M that satisfies $M\mathbb{Z}^2 = 2\mathbb{Z} \times \mathbb{Z}$. Corollary 5.3 and Lemma 5.4 explain how to apply Theorem 2.1 to other dilation matrices, such as $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$.

The statement of Theorem 2.1 contains two assumptions: $B(0) \neq 0$ and $A(1) \neq 0$. Now we explain why these assumptions impose no loss of generality.

First, by shifting the variables in the dilation equation (2.1), one can check that a shift of the coefficient mask by a vector $\mathbf{s}$ results in a shift of the scaling function by $(M - I)\mathbf{s}$. Therefore, without loss of generality, the coefficients of any mask $C(z, w)$ can always be shifted so that $C(z, w)$ contains no negative powers of w, and so that the smallest power of z in $B(z)$ is zero. (The Laurent polynomial $A(z)$ may contain negative powers of z; nevertheless, Theorem 2.1 holds.)

Second, suppose that $A(1) = 0$ and $B(1) \neq 0$ (recall that $A(1) + B(1) = 1$). Change the variables $\mathbf{x} \mapsto -\mathbf{x}$ and $\mathbf{n} \mapsto -\mathbf{n}$ in (2.1), and note (cf. Lemma 5.2) that if $\phi(\mathbf{x})$ solves (2.1) with the coefficient mask $C(z, w)$, then $\phi(-\mathbf{x})$ solves (2.1) with the coefficient mask $\tilde{C}(z, w) = C(z^{-1}, w^{-1})$. Shift the coefficients of $\tilde{C}(z, w)$ to the first quadrant (that is, multiply by $z^N w$, where $N = \max(\deg A, \deg B)$) and observe that $z^N w\, \tilde{C}(z, w) = \tilde{A}(z) + w\tilde{B}(z)$, where $\tilde{A}(1) = B(1) \neq 0$. The shift of $\phi(-\mathbf{x})$ is orthogonal and has accuracy r if and only if $\phi(\mathbf{x})$ is orthogonal and has accuracy r. So the assumption $A(1) \neq 0$ causes no loss of generality in Theorem 2.1.

4. Smoothness. In this section we prove Theorem 2.2. To prove the orthogonality of a scaling function $\phi \in \Phi_{r+1}$, we check Cohen's criterion (4.6). We establish the smoothness of ϕ by estimating the decay of its Fourier transform $\widehat{\phi}$ in a series of lemmas. In this section, $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ and $W = M^{\mathrm{T}}$.

The *symbol* of the dilation equation (2.1) with coefficient mask $C(z, w)$ is the trigonometric polynomial $m(\omega_1, \omega_2) := C(e^{i\omega_1}, e^{i\omega_2})$. Taking the Fourier transform of (2.1) and iterating, we obtain

$$\widehat{\phi}(\omega) = \widehat{\phi}(0) \cdot \prod_{j=1}^{\infty} m(W^{-j}\omega). \tag{4.1}$$

Since $W^{-2} = \frac{1}{2}I$, the infinite product (4.1) breaks into two parts:

$$\prod_{j=1}^{\infty} m(W^{-j}\omega) = \left(\prod_{j=1}^{\infty} m(2^{-j}\omega)\right)\left(\prod_{j=1}^{\infty} m(2^{-j}W\omega)\right). \tag{4.2}$$

Our goal is to derive an estimate for $\prod_{j=1}^{\infty} m(2^{-j}\omega)$ when $C(z, w) \in \mathcal{C}_{r+1}$ by using only the first frequency ω_1.

First, let $C(z, w) = H(z) = \frac{1}{2}(1 + z)$ in (2.1). It can be checked that (2.1) is solved by the Haar scaling function $\chi_{[0,1)^2}$ and that

$$\prod_{j=1}^{\infty} m_H(W^{-j}\omega) = \widehat{\chi_{[0,1)}}(\omega_1) \cdot \widehat{\chi_{[0,1)}}(\omega_2) = \frac{(e^{i\omega_1} - 1)}{i\omega_1} \cdot \frac{(e^{i\omega_2} - 1)}{i\omega_2},$$

where $m_H(\omega_1, \omega_2) := H(e^{i\omega_1})$. Therefore, there is a constant $c > 0$ such that

$$\left|\prod_{j=1}^{\infty} m_H(W^{-j}\omega)\right| \leq c \cdot \big(1 + |\omega_1|\big)^{-1}\big(1 + |\omega_2|\big)^{-1}. \tag{4.3}$$

Next, fix $r \in \mathbb{N}$, and consider any coefficient mask $C(z,w) \in \mathcal{C}_{r+1}$, as described in Theorem 2.1 and (2.11). The corresponding symbol has the form

$$m(\omega_1,\omega_2) = C(e^{i\omega_1}, e^{i\omega_2}) = H^r(e^{i\omega_1})\big(A_0(e^{i\omega_1}) + e^{i\omega_2}B_0(e^{i\omega_1})\big).$$

Define

$$q(\omega_1,\omega_2) := A_0(e^{i\omega_1}) + e^{i\omega_2}B_0(e^{i\omega_1}) \quad \text{and} \quad \ell(\omega_1) := |A_0(e^{i\omega_1})| + |B_0(e^{i\omega_1})|. \tag{4.4}$$

Clearly, $|q(\omega_1,\omega_2)| \le \ell(\omega_1)$. We estimate $\ell(\omega_1)$ in a series of lemmas. Define

$$T_r(y) := \sum_{j=0}^{r-1} \binom{r+j-1}{j} y^j. \tag{4.5}$$

LEMMA 4.1. *Let $L(z)$ satisfy* (2.9). *Assume $r \ge 1$. Then*

$$\mathcal{P}_H^r(-e^{i\omega_1})L^2(-1) \le T_r\left(\sin^2\frac{\omega_1}{2}\right),$$

and equality holds only if $\omega_1 \equiv \pi \pmod{2\pi}$.

Proof. From (2.9) we obtain $L^2(-1) = \mathcal{P}_L(-1) = T_r(1) = \sum_{j=0}^{r-1}\binom{r+j-1}{j}$. By the definition of H in (2.5),

$$\mathcal{P}_H^r(-e^{i\omega_1}) = \left(\frac{1-\cos\omega_1}{2}\right)^r = \left(\sin\frac{\omega_1}{2}\right)^{2r} \le \left(\sin\frac{\omega_1}{2}\right)^{2j} \quad \text{for all } j \le r.$$

Hence

$$\begin{aligned} L^2(-1)\mathcal{P}_H^r(-e^{i\omega_1}) &= \sum_{j=0}^{r-1}\binom{r+j-1}{j}\left(\sin\frac{\omega_1}{2}\right)^{2r} \\ &\le \sum_{j=0}^{r-1}\binom{r+j-1}{j}\left(\sin\frac{\omega_1}{2}\right)^{2j} = T_r\left(\sin^2\frac{\omega_1}{2}\right). \end{aligned}$$

Because the coefficients of T_r are positive, equality holds only when $\sin^2\frac{\omega_1}{2} = 1$, that is, when $\omega_1 \equiv \pi \pmod{2\pi}$. ∎

COROLLARY 4.2. *There exist real polynomials S and L that satisfy* (2.8)–(2.11).

Proof. We need only show that (2.8) defines a valid autocorrelation; that is,

$$1 - \mathcal{P}_H^r(z)\mathcal{P}_H^r(-z)L^2(-1) \ge 0 \quad \text{for all} \quad |z| = 1.$$

Let $|z| = 1$. From Lemma 4.1 and (3.6) we obtain

$$\mathcal{P}_H^r(z)\mathcal{P}_H^r(-z)L^2(-1) \le \mathcal{P}_H^r(z)\mathcal{P}_L(z) = 1 - \mathcal{P}_H^r(-z)\mathcal{P}_L(-z) \le 1.$$

The last inequality holds because $\mathcal{P}_H(z) \ge 0$ and $\mathcal{P}_L(z) \ge 0$ for $|z| = 1$. ∎

LEMMA 4.3. $|A_0(e^{i\omega_1})|^2 \le T_r(\sin^2\frac{\omega_1}{2})$ *and* $|B_0(e^{i\omega_1})|^2 \le T_r(\sin^2\frac{\omega_1}{2})$. *Each equality holds only if $\omega_1 \equiv \pi \pmod{2\pi}$.*

Proof. By the construction in Theorem 2.1 and (2.11) we have

$$\begin{aligned} \mathcal{P}_{A_0}(z) &= \mathcal{P}_L(z)\mathcal{P}_S(z^2) = \mathcal{P}_L(z)\big(1 - \mathcal{P}_H^r(z)\mathcal{P}_H^r(-z)L^2(-1)\big), \\ \mathcal{P}_{B_0}(z) &= \mathcal{P}_H^r(-z)\mathcal{P}_H^r(-z)\mathcal{P}_L(-z)L^2(-1). \end{aligned}$$

Since $\mathcal{P}_H^r(e^{i\omega}) \geq 0$ for all ω,

$$|A_0(e^{i\omega_1})|^2 = \mathcal{P}_{A_0}(e^{i\omega_1}) \leq \mathcal{P}_L(e^{i\omega_1}) \cdot 1 = T_r(\sin^2 \tfrac{\omega_1}{2}).$$

Observe that (3.6) yields $0 \leq \mathcal{P}_H^r(-z)\mathcal{P}_L(-z) \leq 1$ for $z = e^{i\omega_1}$. Therefore,

$$|B_0(e^{i\omega_1})|^2 = \mathcal{P}_{B_0}(e^{i\omega_1}) \leq \mathcal{P}_H^r(-e^{i\omega_1}) \cdot 1 \cdot L^2(-1).$$

Lemma 4.1 completes the inequality for B_0. □

Before we proceed with the estimate of the infinite product (4.2), we recall the following lemma (without proof) from Daubechies [8, pp. 220–226]:

LEMMA 4.4 (Cohen and Conze [5]). *Let* $\mu_r := \frac{1}{2}\log_2 T_r(\frac{3}{4})$. *Then*

$$T_r\big(\sin^2 \tfrac{\omega_1}{2}\big) \leq T_r\big(\tfrac{3}{4}\big) \quad \text{if } |\omega_1| \leq \tfrac{2\pi}{3},$$
$$T_r(\sin^2 \omega_1) T_r\big(\sin^2 \tfrac{\omega_1}{2}\big) \leq T_r^2\big(\tfrac{3}{4}\big) \quad \text{if } \tfrac{2\pi}{3} < |\omega_1| \leq \pi.$$

Furthermore,

$$r - \mu_r > \left(1 - \frac{1}{2}\log_2 3\right) r + \frac{1}{2}\log_2 3,$$
$$\lim_{r\to\infty} \frac{r - \mu_r - 2}{r} = 1 - \frac{1}{2}\log_2 3 \approx 0.2075. \quad \square$$

LEMMA 4.5. *There exist constants* $c > 0$ *and* $\epsilon > 0$ *such that*

$$\prod_{j=1}^{\infty} \ell(2^{-j}\omega_1) \leq c \cdot |\omega_1|^{\mu_r + 1 - \epsilon}, \quad \text{where } \mu_r := \tfrac{1}{2}\log_2 T_r(\tfrac{3}{4}).$$

Proof. Lemma 4.3 implies that $\ell^2(\omega_1) = \big(|A_0(e^{i\omega_1})| + |B_0(e^{i\omega_1})|\big)^2 \leq 4\, T_r(\sin^2 \frac{\omega_1}{2})$ and that equality holds only when $\omega_1 \equiv \pi \pmod{2\pi}$. By combining these statements with Lemma 4.4, we obtain the following strict inequalities:

$$\ell^2(\omega_1) < 4\, T_r\big(\tfrac{3}{4}\big) \quad \text{if } |\omega_1| \leq \tfrac{2\pi}{3},$$
$$\ell^2(\omega_1)\ell^2(2\omega_1) < 4^2\, T_r^2\big(\tfrac{3}{4}\big) \quad \text{if } \tfrac{2\pi}{3} < |\omega_1| \leq \pi.$$

Since ℓ and T_r are continuous, we can preserve these inequalities even after we replace 4 by $(4-\delta)$ for some small $\delta > 0$. By Theorem 2.3 of Cohen and Daubechies [6, p. 69], there exists a constant $c > 0$ such that

$$\prod_{j=1}^{\infty} \ell(2^{-j}\omega_1) \leq c \cdot |\omega_1|^b,$$

where $b := \frac{1}{2}\log_2\big((4-\delta)T_r(\frac{3}{4})\big)$. Observe that $b = 1 - \epsilon + \mu_r$ for some $\epsilon > 0$. □

COROLLARY 4.6. *Let* $W = M^{\mathrm{T}}$. *There exist constants* $c > 0$ *and* $\epsilon > 0$ *such that*

$$\left|\prod_{j=1}^{\infty} q(W^{-j}\omega)\right| \leq c \cdot |\omega_1|^{\mu_r + 1 - \epsilon}\, |\omega_2|^{\mu_r + 1 - \epsilon},$$

where $\mu_r := \frac{1}{2}\log_2 T_r(\frac{3}{4})$, *and* $q(\omega)$ *is defined in* (4.4).

Proof. Note that $W(\omega_1, \omega_2) = (\omega_2, 2\omega_1)$. Therefore, (4.2) and (4.4) yield

$$\begin{aligned}\left|\prod_{j=1}^{\infty} q(W^{-j}\omega)\right| &= \prod_{j=1}^{\infty} \left|q(2^{-j}\omega)\right| \prod_{j=1}^{\infty} \left|q(2^{-j}W\omega)\right| \\ &\le \prod_{j=1}^{\infty} \ell(2^{-j}\omega_1) \prod_{j=1}^{\infty} \ell(2^{-j}\omega_2) \\ &\le c \cdot |\omega_1|^{\mu_r+1-\epsilon} |\omega_2|^{\mu_r+1-\epsilon}. \qquad \square\end{aligned}$$

The following lemma establishes a simple fact about the zeros of $m(\omega_1, \omega_2)$:

LEMMA 4.7. *Suppose that* $m(\omega_1, \omega_2) = 0$. *Then* $\omega_1 \equiv \pi \pmod{2\pi}$.

Proof. If $m(\omega_1, \omega_2) = 0$, then $|A(e^{i\omega_1})| = |B(e^{i\omega_1})|$, and therefore $\mathcal{P}_A(z) = \mathcal{P}_B(z)$ for some $z = e^{i\omega_1}$. Thus, either $H(z) = 0$, and therefore $z = -1$, or $\mathcal{P}_{A_0}(z) = \mathcal{P}_{B_0}(z)$; that is,

$$\left(1 - \mathcal{P}_H^r(-z)\mathcal{P}_H^r(z)L^2(-1)\right)\mathcal{P}_L(z) = \mathcal{P}_H^{2r}(-z)\mathcal{P}_L(-z)L^2(-1),$$

and therefore

$$\mathcal{P}_L(z) = \mathcal{P}_H^r(-z)\mathcal{P}_H^r(z)\mathcal{P}_L(z)L^2(-1) + \mathcal{P}_H^r(-z)\mathcal{P}_H^r(-z)\mathcal{P}_L(-z)L^2(-1).$$

Due to (3.6),

$$\mathcal{P}_L(z) = \mathcal{P}_H^r(-z)L^2(-1).$$

By Lemma 4.1, this can happen only when $\omega_1 \equiv \pi \pmod{2\pi}$. $\square$

Proof of Theorem 2.2.

(i) Since the conditions for accuracy (3.2)–(3.3) and orthogonality (2.3) hold by Theorem 2.1, to ensure that ϕ is an orthogonal scaling function, we need only verify Cohen's criterion by finding a compact fundamental domain Ω of the lattice $2\pi\mathbb{Z}^2$ with the property

$$m(W^{-j}\omega) \neq 0 \quad \text{for all } j \ge 1 \text{ and for all } \omega \in \Omega. \tag{4.6}$$

Let $\Omega = [-\pi, \pi]^2$. Clearly, Ω is a compact fundamental domain of $2\pi\mathbb{Z}^2$. Observe that $W^{-1}\Omega = [-\frac{\pi}{2}, \frac{\pi}{2}] \times [-\pi, \pi]$. Therefore, Lemma 4.7 guarantees that $m(\omega) \neq 0$ for $\omega \in W^{-1}\Omega$. Finally, $W^{-j-1}\Omega \subset W^{-j}\Omega$ for all $j \ge 1$, which proves (4.6).

(ii) Berger and Wang [3] showed that the support of the solution to the dilation equation (2.1) is the attractor of the iterated function system (IFS)

$$\left\{ f_{\mathbf{n}}(\mathbf{x}) := M^{-1}(\mathbf{x} + \mathbf{n}) \ : \ c_{\mathbf{n}} \neq 0 \right\}.$$

If $\operatorname{supp} c = [0, N_x] \times [0, N_y]$, it can be checked that $\operatorname{supp} \phi \subseteq [0, N_x + 2N_y] \times [0, N_x + N_y]$. For all ϕ in Φ_{r+1}, we have $N_x = 4r - 1$ and $N_y = 1$, which proves part (ii).

(iii) Define $m_1(\omega) := m(\omega)\, m(W^{-1}\omega)$. By (4.1), we have $\widehat{\phi}(\omega) = m_1(\omega/2)\, \widehat{\phi}(\omega/2)$. Assume that $\phi(\mathbf{x})$ is separable. Then $\widehat{\phi}(\omega)$ and $m_1(\omega/2)$ are separable. This is impossible, because the coefficient mask $C(z, w)$ is nonseparable by (2.6)–(2.7).

(iv) Let $\mu_r := \frac{1}{2} \log_2 T_r(\frac{3}{4})$. Combining (4.1)–(4.3) with Corollary 4.6, we obtain the following estimate for the decay of the Fourier transform of any ϕ in Φ_{r+1}:

$$|\widehat{\phi}(\omega)| \le c \cdot (1 + |\omega_1|)^{\mu_r+1-r-\epsilon} (1 + |\omega_2|)^{\mu_r+1-r-\epsilon}$$

for some constants $c > 0$ and $\epsilon > 0$. As a result, $\phi \in C^{r-\mu_r-2}(\mathbb{R}^2)$. Lemma 4.4 provides the asymptotics for $r - \mu_r$. ∎

Remark. Observe that, aside from the claim on the size of $\operatorname{supp}\phi$ in part (ii), in the proof of Theorem 2.2 and Corollary 4.6 we used only the following properties of the dilation matrix $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$:

$$M^2 = \pm 2I \quad \text{and} \quad M\mathbb{Z}^2 = 2\mathbb{Z} \times \mathbb{Z}. \tag{4.7}$$

Therefore, aside from part (ii), Theorem 2.2 holds for any dilation matrix M satisfying (4.7), such as $M = \left(\begin{smallmatrix} 0 & -2 \\ 1 & 0 \end{smallmatrix}\right)$. Lemma 5.2 and a result by Lagarias and Wang [11] extend Theorem 2.2 to all dilation matrices with $M^2 = \pm 2I$ [2].

5. Examples. In this section we consider a few of the coefficient masks in Theorem 2.1 for $r = 1$, 2, and 6. We plot the corresponding scaling functions and study their smoothness. Finally, we discuss how to modify these masks for other dilation matrices. Additional coefficient values can be found in the appendix of [2].

Devil's Tower. The simplest case of Theorem 2.1 is when $r = 1$ and (2.11) holds. In this case $L(z) = 1$, and the autocorrelation

$$\mathcal{P}_S(z^2) = 1 - \Big(\frac{1 - z^2}{4}\Big)\Big(\frac{1 - z^{-2}}{4}\Big)$$

has only two spectral factors, which can be computed explicitly:

$$S^{(1)}(z^2) = \frac{2 - \sqrt{3}}{4} + \frac{2 + \sqrt{3}}{4} z^2 \quad \text{and} \quad S^{(2)}(z^2) = \frac{2 + \sqrt{3}}{4} + \frac{2 - \sqrt{3}}{4} z^2.$$

The corresponding coefficient masks are given below. For brevity we omit the powers of z and w; the constant term c_{00} is anchored at the lower-left corner:

$$c^{(1)} = \frac{1}{8} \begin{bmatrix} -1 & 1 & 1 & -1 \\ 2 - \sqrt{3} & 2 - \sqrt{3} & 2 + \sqrt{3} & 2 + \sqrt{3} \end{bmatrix},$$

$$c^{(2)} = \frac{1}{8} \begin{bmatrix} -1 & 1 & 1 & -1 \\ 2 + \sqrt{3} & 2 + \sqrt{3} & 2 - \sqrt{3} & 2 - \sqrt{3} \end{bmatrix}.$$

The corresponding scaling functions $\phi^{(1)}$ and $\phi^{(2)}$ are supported on $[0, 5] \times [0, 4]$, have accuracy 2, and are discontinuous. The mesh plot of $\phi^{(1)}$ resembles the famous Wyoming mountain, as depicted in Figure 5.1.

Continuous scaling function. In the case $r = 2$, four spectral factors of $\mathcal{P}_S$ are combined with two spectral factors of $\mathcal{P}_L$ to produce eight coefficient masks in $\mathcal{C}_{2+1}$ (see Table A.1 in the Appendix). The second mask has the following coefficients:

$$c = \begin{bmatrix} 0.03697 & -0.06403 & -0.05678 & 0.13797 & 0.00265 & -0.08384 & 0.01716 & 0.00991 \\ 0.00790 & -0.01369 & -0.13808 & 0.12118 & 0.54993 & 0.34617 & 0.08025 & 0.04633 \end{bmatrix}. \tag{5.1}$$

The corresponding scaling function (Figure 5.2) has accuracy $2 + 1 = 3$ and is continuous. The continuity does not follow from Theorem 2.2, because $2 - \mu_2 - 2 \approx 1.339 - 2 < 0$. Instead, the continuity of the scaling function in Figure 5.2 follows from Lemma 5.1, which is a straightforward modification of Lemma 7.1.2 by Daubechies [8, p. 217]; the proof is essentially the same and is therefore omitted here.

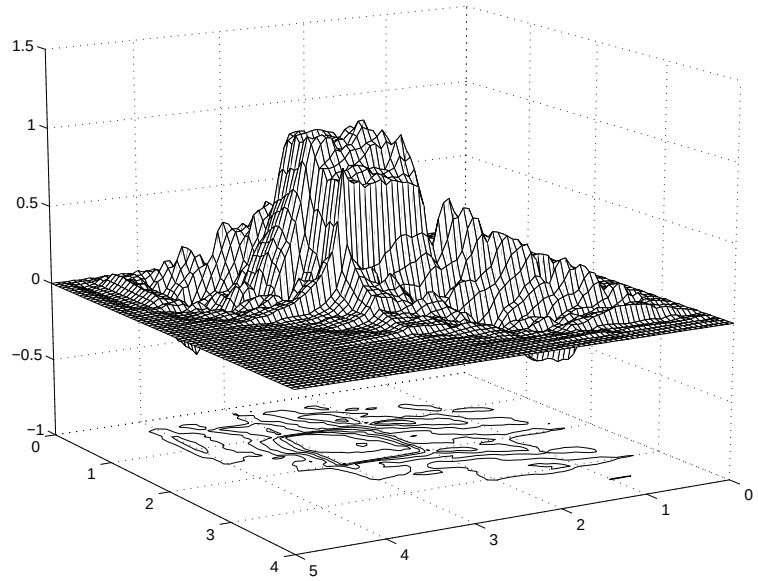


FIG. 5.1. *"Devil's Tower": a discontinuous scaling function with accuracy* $1+1=2$ *and support* $[0,5]\times[0,4]$.

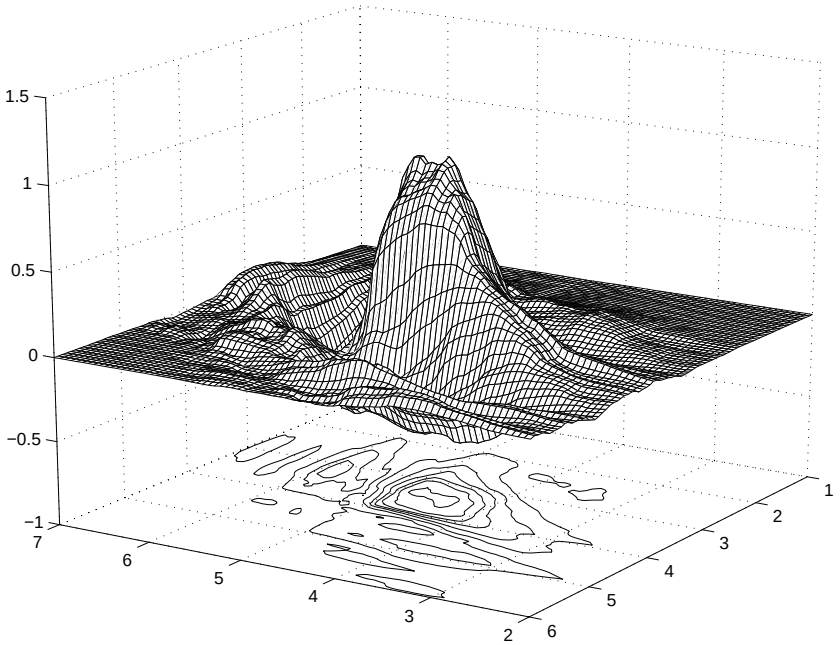


FIG. 5.2. *"Resting Dog": a continuous nonseparable scaling function with accuracy* $2+1=3$ *and support* $[0,9]\times[0,8]$.

LEMMA 5.1. *Let $W = M^{\mathrm{T}}$. Suppose that there exist $p \in \mathbb{N}$ and $\lambda > 0$ such that*

$$\sup_{\omega} \prod_{j=0}^{p-1} \left|q(2^{-j}\omega)\, q(2^{-j}W^{-1}\omega)\right| < 2^{\lambda p}, \tag{5.2}$$

where $q(\omega)$ is defined in (4.4). *Then there exist constants c' and c'' such that*

$$\prod_{j=1}^{\infty} \left|q(W^{-j}\omega)\right| < c' \cdot (1+|\omega|)^{\lambda} < c'' \cdot (1+|\omega_1|)^{\lambda}(1+|\omega_2|)^{\lambda}.$$

Hence $\phi \in C^{r-1-\lambda}$. ∎

The authors evaluated the trigonometric polynomial in (5.2) numerically for 120 values of ω within a single period and verified that the coefficient mask (**??**) satisfies (5.2) for $p = 4$ and $\lambda = 1$. Therefore, the scaling function in Figure 5.2 is continuous.

While Theorem 2.2 claims only that $\Phi_{5+1} \subset C^0$, it can be verified numerically that all masks in $\mathcal{C}_{3+1}$ (see Table A.3 in the Appendix) satisfy (5.2) for $p = 4$ and $\lambda = 2$; hence $\Phi_{3+1} \subset C^0$.

Differentiable scaling functions. About half of the 128 coefficient masks in $\mathcal{C}_{6+1}$ satisfy (5.2) for $p = 4$ and $\lambda = 4$ and therefore produce differentiable scaling functions. To the best of the authors' knowledge, these are the first nonseparable orthogonal scaling functions (for a dilation matrix with $|\det(M)| = 2$) that are differentiable. The 2×24 scaling coefficients of the differentiable scaling function in Figure 5.3 are displayed in Table A.2 in the Appendix. Numerical computations confirm also that $\Phi_{7+1} \subset C^1$.

Other dilations. The following lemma explains what happens when we manipulate the coefficient mask in the dilation equation (2.1). Let J be any lattice-preserving matrix, that is, any matrix J such that both J and J^{-1} have integer entries. An example is $J = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$. For any function $f : \mathbb{R}^2 \to \mathbb{R}$ define its "upsampling by J" as $(J \uparrow f)(\mathbf{x}) := f(J^{-1}\mathbf{x})$. Similarly, define the "upsampled by J" scaling coefficients as $(J \uparrow c)_{\mathbf{n}} := c_{J^{-1}\mathbf{n}}$.

LEMMA 5.2. *Let ϕ be a solution to the dilation equation* (2.1) *with coefficients c and dilation matrix M. Then $J \uparrow \phi$ is a solution to the dilation equation* (2.1) *with coefficients $J \uparrow c$ and dilation matrix JMJ^{-1}.*

Proof. Make the substitution $\mathbf{x} \mapsto J^{-1}\mathbf{y}$ in the dilation equation. ∎

A surprising example is $J = \left(\begin{smallmatrix} -1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$, the matrix of the reflection about the y-axis. Let $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$. Then $JMJ^{-1} = -M$. Therefore, the scaling function with coefficient mask $w^{-1}(B(z) + wA(z))$ and dilation M is a reflection of the scaling function with coefficient mask $A(z) + wB(z)$ and dilation $-M$, but it is *not* a reflection of the scaling function with coefficient mask $A(z) + wB(z)$ and dilation M. The difference is obvious in Figure 5.4, which shows the contour plots of two scaling functions with the same dilation matrix $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ and x-reflected coefficients. (The one on the left was plotted in Figure 5.2.)

The plots exhibit a peculiar feature: all scaling functions appear so far to be "almost symmetric" with respect to the bisectrix $x = y$. This is not surprising, since the dilation matrix $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ swaps the axes and stretches along x by a factor of two. Unfortunately, the case $|\det(M)| = 2$ is similar to the univariate case $M = 2$—the only symmetric orthogonal scaling function is the Haar function.

Lemma 5.2 allows us to adapt the coefficient masks described in Theorem 2.1 to dilation matrices that generate the quincunx sublattice in the following way.

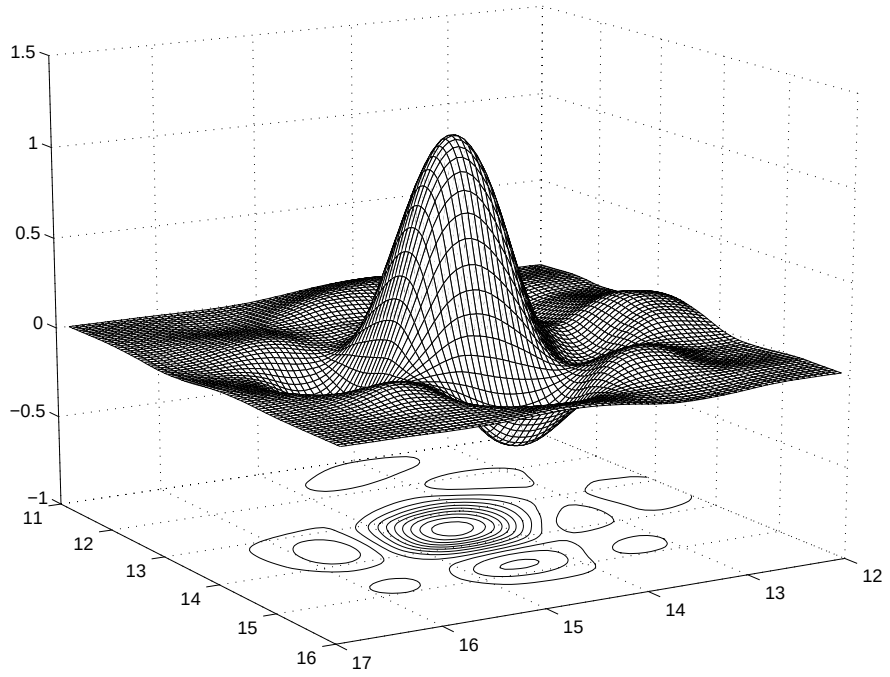


FIG. 5.3. *A differentiable nonseparable scaling function with accuracy* $6 + 1 = 7$.

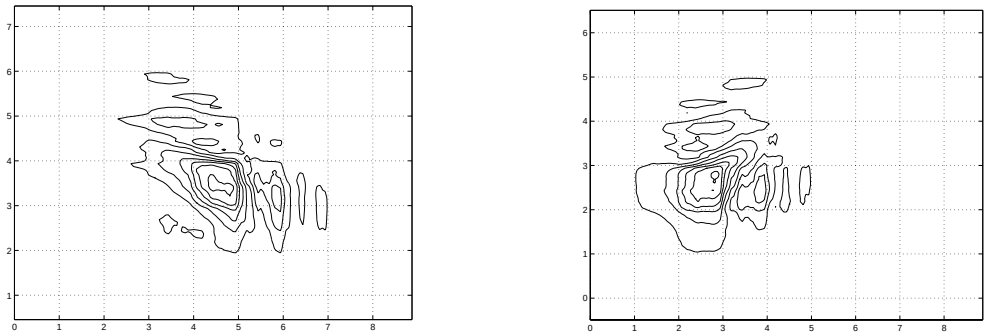


FIG. 5.4. *Level sets of scaling functions with x-reflected scaling coefficients: the functions are not similar.*

COROLLARY 5.3. *Let* ν *be odd. If* $A(z) + wB(z)$ *is a two-row orthogonal coefficient mask with accuracy r for the column lattice, then* $A(z) + z^\nu wB(z)$ *is a two-row orthogonal coefficient mask with accuracy r for the quincunx lattice.*

Proof. Set $M = \left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ and $J = \left(\begin{smallmatrix} 1 & \nu \\ 0 & 1 \end{smallmatrix}\right)$ in Lemma 5.2. Observe that the J-upsampling of $A(z) + wB(z)$ is $A(z) + z^\nu wB(z)$ and that the matrix $J\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)J^{-1}$ generates the quincunx sublattice. □

If we shift the top row of coefficients in (**??**) one position to the left and use the dilation matrix $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ in (2.1), then we obtain an orthogonal nonseparable quincunx scaling function with accuracy $2 + 1 = 3$, plotted in Figure 5.5 (cf. Kovačević and Vetterli's scaling function with accuracy 2 [12]).

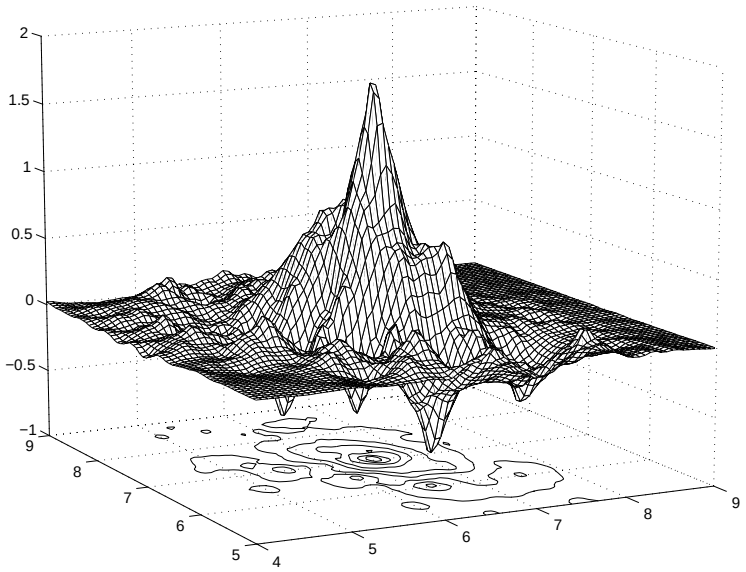


FIG. 5.5. *An orthogonal nonseparable scaling function with accuracy* $2+1=3$ *for the quincunx dilation matrix.*

Every 2×2 dilation matrix M with $M^2 = 2I$ can be factored in the form $J\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)J^{-1}$ [11]. Therefore, Lemma 5.2 and Theorem 2.2 provide arbitrarily smooth nonseparable wavelet bases for such dilation matrices.

The convolution of two sets of scaling coefficients $a_{\mathbf{n}}$ and $b_{\mathbf{n}}$ is defined by $(a*b)_{\mathbf{n}} := \sum_{\mathbf{k}\in\mathbb{Z}^2} a_{\mathbf{k}}\, b_{\mathbf{n}-\mathbf{k}}$. The following statement combines two levels of the MRA into one.

LEMMA 5.4. *Let* ϕ *be a solution to the dilation equation* (2.1) *with coefficients* c *and dilation matrix* M*. Then* ϕ *is also a solution to the dilation equation* (2.1) *with coefficients* $c * (M \uparrow c)$ *and dilation matrix* M^2*.*

Proof. Iterate the dilation equation (2.1); that is, replace each ϕ on the right by the sum of its dilates. □

Since $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)^2 = 2I$, by Theorem 2.2 and Lemma 5.4, we obtain the following.

COROLLARY 5.5. *Aside from the separable Daubechies MRA, there exists a bivariate nonseparable orthogonal MRA of any given smoothness for the dilation* $\left(\begin{smallmatrix} 2 & 0 \\ 0 & 2 \end{smallmatrix}\right)$.

Ayache [1] obtained this result independently by perturbing the separable Daubechies basis.

6. Summary. We showed how to obtain orthogonal two-row coefficient masks (low-pass filter coefficients) with arbitrarily high accuracy for dilation matrices M with $|\det(M)| = 2$, such as $\left(\begin{smallmatrix} 0 & 2 \\ 1 & 0 \end{smallmatrix}\right)$ or $\left(\begin{smallmatrix} 1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$. We proved that if $M^2 = \pm 2I$, then the smoothness of the scaling functions corresponding to those coefficient masks increases asymptotically with the accuracy and can be made arbitrarily high.

Appendix. Scaling coefficients. The following tables contain the coefficients of $A(z)$ and $B(z)$ that satisfy the conditions (2.6)–(2.11) in Theorem 2.1, that is, the coefficients of $C(z,w)$ in $\mathcal{C}_{r+1}$ for some values of r. Generally speaking, the scaling coefficients in the beginning of each table correspond to the "least symmetric" scaling functions; the scaling coefficients towards the end of the table correspond to the "most symmetric" scaling functions.

TABLE A.1
Scaling coefficients (first half of $\mathcal{C}_{2+1}$) of accuracy $2+1$.

n	a_n	b_n	a_n	b_n
	Solution 1		Solution 2	
0	0.001113697631	0.036969146934	0.007903570868	0.036969146934
1	-0.001928980881	-0.064032440802	-0.013689386305	-0.064032440802
2	-0.011710875083	-0.056780853066	-0.138084015734	-0.056780853066
3	0.003658326071	0.137970734671	0.121182418797	0.137970734671
4	-0.058943362439	0.002654265329	0.549926093940	0.002654265329
5	0.169446270789	-0.083844146934	0.346172096397	-0.083844146934
6	0.569540539891	0.017157440802	0.080254350926	0.017157440802
7	0.328824384020	0.009905853066	0.046334871111	0.009905853066
	Solution 3		Solution 4	
0	-0.012415391296	0.036969146934	-0.088108228150	0.036969146934
1	0.021504088520	-0.064032440802	0.152607927721	-0.064032440802
2	-0.006740179593	-0.056780853066	0.565028719464	-0.056780853066
3	0.197013817950	0.137970734671	0.336639086235	0.137970734671
4	0.570245056104	0.002654265329	0.030278563342	0.002654265329
5	0.310978621572	-0.083844146934	0.014909362188	-0.083844146934
6	-0.051089485215	0.017157440802	-0.007199054655	0.017157440802
7	-0.029496528042	0.009905853066	-0.004156376143	0.009905853066

TABLE A.2
The scaling coefficients of the differentiable scaling function in Figure 5.3.

n	a_n	b_n
0	0.000143412339	0.000082104697
1	-0.000323736348	-0.000185341617
2	-0.001920307470	-0.000431083617
3	0.003595546890	0.000549853896
4	0.008522556302	0.003143374464
5	-0.004408987885	-0.002050950686
6	-0.013081025520	-0.013941778594
7	-0.060122740069	0.012068826968
8	-0.098853166909	0.031125316212
9	0.162212221109	-0.036965110063
10	0.525503586261	-0.036331876410
11	0.406747601563	0.060643179987
12	0.037817366952	0.019734738252
13	-0.036595246129	-0.057961647081
14	0.043460344536	0.000210716144
15	0.031894654283	0.033234701933
16	-0.002789322971	-0.006059880986
17	-0.003862049736	-0.011382637768
18	0.001292258468	0.003005793635
19	0.000931477590	0.002333217232
20	-0.000121744397	-0.000582912091
21	-0.000080277824	-0.000304243713
22	0.000026042408	0.000045488295
23	0.000011536556	0.000020150912

TABLE A.3
Scaling coefficients (first half of $\mathcal{C}_{3+1}$) of accuracy $3+1$.

n	a_n	b_n	a_n	b_n
	Solution 1		Solution 2	
0	0.000062409389	-0.011623030789	0.000326746845	-0.011623030789
1	-0.000151373802	0.028191629145	-0.000792523573	0.028191629145
2	-0.000616099021	0.018801633679	-0.004873897732	0.018801633679
3	0.001728930996	-0.089291978349	0.013049802109	-0.089291978349
4	0.003825989061	0.016318478955	0.044932505092	0.016318478955
5	-0.006639890465	0.099956916164	-0.101357239374	0.099956916164
6	0.011989717589	-0.045534931049	-0.168647288127	-0.045534931049
7	-0.050002611258	-0.046035169850	0.262229312212	-0.046035169850
8	-0.071220291676	0.025023044461	0.522072641289	0.025023044461
9	0.325850800478	0.008409358878	0.283090226735	0.008409358878
10	0.555958274659	-0.002985195258	0.106189292633	-0.002985195258
11	0.229214144050	-0.001230755988	0.043780421891	-0.001230755988
	Solution 3		Solution 4	
0	0.004635883479	-0.011623030789	0.024271352757	-0.011623030789
1	-0.011244322602	0.028191629145	-0.058870099219	0.028191629145
2	0.005799855170	0.018801633679	-0.092072378476	0.018801633679
3	0.003357833887	-0.089291978349	0.314552501801	-0.089291978349
4	0.012240306723	0.016318478955	0.551933076802	0.016318478955
5	0.360239746045	0.099956916164	0.242555222384	0.099956916164
6	0.554047592739	-0.045534931049	0.023310939204	-0.045534931049
7	0.174496893426	-0.046035169850	0.003917651739	-0.046035169850
8	-0.084208083428	0.025023044461	-0.008872536334	0.025023044461
9	-0.029935887232	0.008409358878	-0.002744659377	0.008409358878
10	0.007484445316	-0.002985195258	0.001429546047	-0.002985195258
11	0.003085736475	-0.001230755988	0.000589382672	-0.001230755988

Note. To save space, we displayed only the *first half* of $\mathcal{C}_{r+1}$. To obtain the other half, simply take the coefficients (both a_n and b_n) in reverse order. Recall that the scaling functions with scaling coefficients in the second half of the family $\mathcal{C}_{r+1}$ are different than, and are not mere flips of, the scaling functions with scaling coefficients in the first (tabulated) half (see the discussion after Lemma 5.2).

Acknowledgments. The authors thank the referees for their useful suggestions, and Albert Cohen for the helpful discussions and for communicating Ayache's work. Special thanks go to Chris Heil, for providing the preprint [4] of the accuracy condition and for carefully revising the early drafts, and to Tim Flaherty for his meticulous work in discovering several inaccuracies in the manuscript. Thanks to Huibin Zhou for his comments. Wang acknowledges the support of the Mathematics Department at Cornell University, where he was a visitor during the completion of this paper. The fruitful blend of theoretical mathematics and computer numerics was essential to our work; the plots and the numerical computations were aided by Mathematica and MATLAB.

REFERENCES

[1] A. Ayache, *Construction of nonseparable dyadic compactly supported orthonormal wavelet bases for $L^2(\mathbb{R}^2)$ of arbitrarily high regularity*, Rev. Mat. Iberoamericana, to appear.

[2] E. Belogay, *Construction of Smooth Orthogonal Wavelets with Compact Support in $\mathbb{R}^d$*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1998.

[3] M. A. Berger and Y. Wang, *Multidimensional two-scale dilation equations*, Wavelets: A Tutorial in Theory and Applications, Wavelet Anal. Appl. 2, C. K. Chui, ed., Academic Press, New York, (1992), pp. 295–323.

[4] C. Cabrelli, C. Heil, and U. Molter, *Accuracy of lattice translates of several multidimensional refinable functions*, J. Approx. Theory, 95 (1998), pp. 5–52.

[5] A. Cohen and J. P. Conze, *Régularité des bases d'ondelettes et mesures ergodiques*, Rev. Mat. Iberoamericana, 8 (1992), pp. 351–365.

[6] A. Cohen and I. Daubechies, *Nonseparable bidimensional wavelet bases*, Rev. Mat. Iberoamericana, 9 (1993), pp. 51–137.

[7] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 41 (1988), pp. 909–996.

[8] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics 61, SIAM, Philadelphia, PA, 1992.

[9] K. Gröchenig and W. Madych, *Multiresolution analysis, Haar bases, and self-similar tilings,* IEEE Trans. Inform. Theory, 38 (1992), pp. 558–568.

[10] W. He and M. J. Lai, *Examples of bivariate nonseparable compactly supported orthonormal continuous wavelets*, in Wavelet Applications in Signal and Image Processing IV, Proceedings, SPIE 3169, 1997, pp. 303–314.

[11] J. C. Lagarias and Y. Wang, *Haar type orthonormal wavelet basis in $\mathbb{R}^2$*, J. Fourier Anal. Appl., 2 (1995), pp. 1–14.

[12] J. Kovačević and M. Vetterli, *Nonseparable multidimensional perfect reconstruction filter-banks,* IEEE Trans. Inform. Theory, 38 (1992), pp. 533–555.

[13] S. Mallat, *Multiresolution analysis and wavelets*, Trans. Amer. Math. Soc., 315 (1989), pp. 69–88.

[14] Y. Meyer, *Ondelettes et Opérateurs*, Hermann, Paris, 1990.

[15] G. Strang and T. Nguyen, *Wavelets and Filter Banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.

[16] L. Villemoes, *Continuity of nonseparable quincunx wavelets,* Appl. Comput. Harmon. Anal., 1 (1994), pp. 180–187.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 699–720

INVERSE CONDUCTIVITY PROBLEM WITH ONE MEASUREMENT: ERROR ESTIMATES AND APPROXIMATE IDENTIFICATION FOR PERTURBED DISKS*

EUGENE FABES†, HYEONBAE KANG‡, AND JIN KEUN SEO§

Abstract. This paper studies the global uniqueness and stability questions of the inverse conductivity problem to determine the unknown object D entering $\mathrm{div}((1+(k-1)\chi_D)\nabla u)=0$ in Ω and $\frac{\partial u}{\partial \nu}=g$ on $\partial\Omega$ from the boundary measurement $\Lambda_D(g)=u|_{\partial\Omega}$. The results of this paper are fourfold. We first obtain a Hölder stability estimate for disks. Second, a uniform stability estimate for the direct problem is obtained. Third, we obtain the stability estimates $|D_1 \setminus \tilde{D}_2| + |D_2 \setminus \tilde{D}_1| \le C(\|\Lambda_{D_1}(g)-\Lambda_{D_2}(g)\|^{\alpha}_{L^\infty(\partial\Omega)}+\epsilon)$ for some $\alpha>0$ when g satisfies some condition if D_1 and D_2 are ϵ-perturbations of two disks. We then drop the condition on g and show that if $\Lambda_{D_1}(g)=\Lambda_{D_2}(g)$ on $\partial\Omega$, then the two domains must be very close.

Key words. inverse conductivity problems, one measurement, uniqueness, stability

AMS subject classifications. 86A20, 58G10, 31A25

PII. S0036141097324958

1. Introduction.

This paper studies the inverse problem to determine the unknown object D entering the Neumann problem

$$P[D,g]\begin{cases} \mathrm{div}((1+(k-1)\chi_D)\nabla u)=0 \quad \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu}=g \quad \text{on } \partial\Omega, \quad \displaystyle\int_{\partial\Omega} u=0, \quad \int_{\partial\Omega} g=0, \quad g\in L^2(\partial\Omega), \end{cases}$$

from the single Cauchy data $(u|_{\partial\Omega}, g)$. By the uniqueness of the Neumann problem $P[D,g]$, we can define the Neumann-to-Dirichlet map Λ_D by

$$\Lambda_D(g) := u|_{\partial\Omega}, \quad g\in L^2_0(\partial\Omega) := \left\{ \psi\in L^2(\partial\Omega) : \int_{\partial\Omega}\psi=0 \right\},$$

where u is the solution of the Neumann problem $P[D,g]$. We are interested in the uniqueness and stability questions, which we state roughly as follows:

Uniqueness: Does $\Lambda_{D_1}(g)=\Lambda_{D_2}(g)$ imply $D_1=D_2$? $\quad(\overline{D_j}\subset\Omega)$

Stability: If $\|\Lambda_{D_1}(g)-\Lambda_{D_2}(g)\|_{L^\infty(\partial\Omega)}$ is small, is $|D_1\Delta D_2|$ small?

Here $|D_1\Delta D_2| = |D_1\setminus D_2| + |D_2\setminus D_1|$ and $|E|$ denotes the Lebesgue measure of the set E.

This paper is concerned primarily with global uniqueness and stability within the class of small perturbation of disks. To explain our result, let us fix the notion of ϵ-perturbations of disks. Let ϵ be a positive number and let Ω_0 be an open subset of

*Received by the editors July 21, 1997; accepted for publication (in revised form) February 10, 1998; published electronically April 20, 1999. The second and third authors were partially supported by GARC-KOSEF, KOSEF, and BSRI.

http://www.siam.org/journals/sima/30-4/32495.html

†This author is deceased. Former address: School of Mathematics, University of Minnesota, Minneapolis, MN 55455.

‡Department of Mathematics, Seoul National University, Seoul 151-742, Korea (hkang@math.snu.ac.kr).

§Department of Mathematics, Yonsei University, Seoul 120-749, Korea (seoj@bubble.yonsei.ac.kr).

Ω at some distance, say, $2\delta_0$, from $\partial\Omega$. A C^2-domain D being an ϵ-perturbation of a disk B means that there is $\omega \in C^1(\partial B)$ with $\|\omega\|_{C^1(\partial B)} \le 1$ such that

$$\partial D : x + \epsilon\omega(x)\nu(x), \quad x \in \partial B, \tag{1.1}$$

where $\nu(x)$ is the outward unit normal to ∂B at x. Let $\mathcal{C}[\epsilon]$ denote the class of ϵ-perturbations of all disks contained in Ω_0 with the radius larger than a fixed number d_0.

The results of this paper are fourfold. The first two results are of preparatory nature for the last two results. However, they have their own interest.

The first result is the Hölder stability estimate within the class of disks. We prove that if the Neumann data g satisfy the conditions

(N1) there exists a positive number M such that $|g'(P)| > M$ if $|g(P)| < M$, $P \in \partial\Omega$ (here, g' means the tangential derivative on $\partial\Omega$); and

(N2) $\{P \in \partial\Omega : g(P) > 0\}$ and $\{P \in \partial\Omega : g(P) < 0\}$ are nonempty connected subsets of $\partial\Omega$,

then

$$|D_1 \Delta D_2| \le C\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|^{\alpha}_{L^\infty(\partial\Omega)} \tag{1.2}$$

for every disk D_1, D_2 contained in Ω_0 and $0 < \alpha < 1$. (See Theorem 3.1.) The conditions on the Neumann data g guarantee the existence of the lower bound of $|\nabla u|$, which depends only on M when $D \in \mathcal{C}[\epsilon]$ and u is the weak solution to $P[D, g]$. (See [KSS1].) This result is an improvement upon the previous logarithmic estimate in [KSS1].

The second result is a uniform stability for the direct problem; namely, if D_ϵ is an ϵ-perturbation of C^2-domain D (D is not necessarily a disk) and u, u_ϵ are solutions of $P[D, g]$, $P[D_\epsilon, g]$, respectively, then

$$\|u - u_\epsilon\|_{L^\infty(\Omega)} \le C\epsilon\|g\|_{L^2(\partial\Omega)}.$$

(See Theorem 4.1.) This result improves upon the L^2-estimate of [BFI].

The third main result of this paper deals with global stability within the class $\mathcal{C}[\epsilon]$. We prove that if the Neumann data g satisfy the conditions (N1) and (N2), then

$$|D_1 \Delta D_2| \le C\left(\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|^{\alpha}_{L^\infty(\partial\Omega)} + \epsilon\right) \tag{1.3}$$

for every $D_1, D_2 \in \mathcal{C}[\epsilon]$ and for some constants C and $0 < \alpha < 1$ independent of ϵ. (See Theorem 5.2.)

The last main result of this paper is concerned with the global uniqueness without any restriction on the Neumann data g. Let $D_0 \in \mathcal{C}[\epsilon]$ and let $\Lambda_{D_0}(g) = f$. If $D \in \mathcal{C}[\epsilon]$ and $\Lambda_D(g) = \Lambda_{D_0}(g)$, we show that

$$|D_0 \Delta D| \le C\epsilon, \tag{1.4}$$

where the constant C depends on (f, g) not on D or ϵ. The estimate (1.4) means that if the boundary measurements are the same, then the two domains must be very close. For this reason we call this result an approximate identification. It seems that this approximate identification of a domain is quite meaningful in a practical sense. This result together with the local uniqueness result of Alessandrini, Isakov, and Powell [AIP] gives the global uniqueness within $\mathcal{C}[\epsilon]$ provided that ϵ is sufficiently small and g satisfies a certain condition.

The classes of domains for which the global uniqueness is proved so far are only polygons, cylinders, and disks in the plane, convex polyhedra, and balls in three dimensions. (See [BFS, FI, IP, KS1, KS2, S].) The log-type global stability is obtained in [KSS1] within the class of the disks. There are local uniqueness and local stability results in two dimensions (see [AIP] and [BFI]).

The organization of this paper is as follows; in section 2 we review the representation formula in [KS1], and some other preliminary results are reviewed or obtained. In section 3 we prove the Hölder stability for the disks. In section 4 we obtain the uniform stability of the solutions to the direct problem $P[D, g]$ under perturbation of domains. The third main result (1.3) is proved in section 5 and the last one (1.4) in section 6.

2. Notation and preliminary definitions. Let us define the Lipschitz character.

Definition of the Lipschitz character. A bounded open connected domain $D \subset \mathbb{R}^2$ is called a Lipschitz domain with Lipschitz character (r_0, L) if for each $P \in \partial D$ there is an open rectangle $Z(P, r_0)$ centered at P, with the bottom length equal to r_0, whose bottom and top sides are at a positive distance l, $r_0 < l < 2r_0$ from ∂D, such that there is a coordinate system $(t, s) \in \mathbb{R} \times \mathbb{R}$, with the s-axis containing the axis of Z and a Lipschitz function $\phi : \mathbb{R} \to \mathbb{R}$ such that $Z \cap D = \{(t, s) \in \mathbb{R} \times \mathbb{R} : s > \phi(t)\} \cap Z$, $Z \cap \partial D = \{(t, s) \in \mathbb{R} \times \mathbb{R} : s = \phi(t)\} \cap Z$, and $\|\phi'\|_{L^\infty(\mathbb{R})} \leq L$.

Throughout this paper, we assume that Ω and D are simply connected bounded Lipschitz domains in $\mathbb{R}^2$ with Lipschitz character (r_0, L) and

$$D \subset \Omega_0 := \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) > 2\delta_0\},$$

where δ_0 is a fixed positive number. We denote by $B_d(a)$ the disk centered at a with radius d.

Let

$$\begin{aligned}
\mathcal{S}_\Omega \phi(x) &= \frac{1}{2\pi} \int_{\partial\Omega} \log|x - y| \phi(y) d\sigma_y, \quad x \in \mathbb{R}^2, \\
\mathcal{D}_\Omega \phi(x) &= \frac{1}{2\pi} \int_{\partial\Omega} \frac{\langle x - y, \nu_y \rangle}{|x - y|^2} \phi(y) d\sigma_y, \quad x \in \mathbb{R}^2 \setminus \partial\Omega, \\
\mathcal{K}_D \phi(x) &= \frac{1}{2\pi} \int_{\partial D} \frac{\langle x - y, \nu_y \rangle}{|x - y|^2} \phi(y) d\sigma_y, \quad x \in \partial D,
\end{aligned}$$

and let $\mathcal{K}_D^*$ be the dual of $\mathcal{K}_D$. Recall the classical trace formula (see [F] or [V])

$$\lim_{t \to 0^+} \langle \nu_P, \nabla S_D \phi(P \pm t\nu_P) \rangle = \left(\pm \frac{1}{2} I + \mathcal{K}_D^* \right) \phi(P) \quad \text{almost all } P \in \partial D. \tag{2.1}$$

From [KS1] and [KSS2], the weak solution u to the Neumann problem $P[D, g]$ can be uniquely expressed as

$$u(x) = H(x) + \mathcal{S}_D \varphi_D(x) \quad \text{for } x \in \Omega, \tag{2.2}$$

where

$$H(x) = -\mathcal{S}_\Omega g(x) + \mathcal{D}_\Omega f(x), \quad f = u|_{\partial\Omega}, \quad x \in \mathbb{R}^2 \setminus \partial\Omega, \tag{2.3}$$

$$(\lambda I - \mathcal{K}_D^*) \varphi_D = \frac{\partial H}{\partial \nu}|_{\partial D} \quad \text{on } \partial D, \tag{2.4}$$

where $\lambda = \frac{k+1}{2(k-1)}$. Moreover, we have

$$\int_{\partial D} \varphi_D = 0, \tag{2.5}$$

$$H(x) + \mathcal{S}_D \varphi_D(x) = 0, \quad x \in \mathbb{R}^2 \setminus \overline{\Omega}. \tag{2.6}$$

Let $u^i := u|_D$ and $u^e := u|_{\Omega \setminus \overline{D}}$. Then

$$\varphi_D = (k-1)\frac{\partial u^i}{\partial \nu} = \frac{k-1}{k}\frac{\partial u^e}{\partial \nu} \quad \text{on } \partial D$$

and hence

$$\varphi_D = \frac{\partial u^e}{\partial \nu} - \frac{\partial u^i}{\partial \nu} \quad \text{on } \partial D. \tag{2.7}$$

(See [KS1] and [KSS2] for proofs.)

In [EFV], it was shown that there is a positive constant C depending only on the Lipschitz character (r_0, L) of D and δ_0 so that

$$\|(\nabla u)^{**}\|_{L^2(\partial D)} \le C \|\nabla u\|_{L(\Omega)}, \tag{2.8}$$

where $(\nabla u)^{**}(x)$ is the interior and exterior nontangential maximal function of ∇u at $x \in \partial D$:

$$(\nabla u)^{**}(x) = \sup\{|\nabla u(y)| : |x-y| \le (2L+1)\mathrm{dist}(y, \partial D),\ y \in \Omega_0\}.$$

Because of the above regularity estimate (2.8), the following transmission condition holds in the $L^2(\partial D)$-sense:

$$\frac{\partial u^e}{\partial \nu} = k \frac{\partial u^i}{\partial \nu} \quad \text{on } \partial D. \tag{2.9}$$

LEMMA 2.1. *Let u be the solution of the Neumann problem $P[D, g]$ and H be as in* (2.3). *Then there is a positive constant C depending on the Lipschitz character of Ω and δ_0 (independent of D) so that*

$$\int_{\partial\Omega} |\nabla u|^2 + \int_\Omega |\nabla u|^2 + |\nabla H|^2 \le C \int_{\partial\Omega} |g|^2. \tag{2.10}$$

Proof. The proof of this lemma is based on the Rellich identity. Let $\vec{\alpha}$ be the smooth vector field such that the support of $\vec{\alpha}$ lies in $\mathbb{R}^2 \setminus \Omega_0$ and $\langle \vec{\alpha}, \nu \rangle > c_1 > 0$ on $\partial\Omega$ (here, c_1 depend on the Lipschitz character of $\partial\Omega$). If $v \in C^2(\overline{\Omega})$, we have

$$\mathrm{div}(\vec{\alpha}|\nabla v|^2) = 2\mathrm{div}(\nabla v \langle \vec{\alpha}, \nabla v\rangle) + \mathrm{div}\vec{\alpha}|\nabla v|^2 - 2\langle \nabla\vec{\alpha}\nabla v, \nabla v\rangle - 2\langle \vec{\alpha}, \nabla v\rangle \Delta v$$

and by integrating both sides of the above identity over Ω we obtain the Rellich identity

$$\int_{\partial\Omega} \langle \vec{\alpha}, \nu\rangle \left|\frac{\partial v}{\partial \nu}\right|^2 = \int_{\partial\Omega} \langle \vec{\alpha}, \nu\rangle \left|\frac{\partial v}{\partial T}\right|^2 - 2\int_{\partial\Omega} \langle \vec{\alpha}, T\rangle \frac{\partial v}{\partial T}\frac{\partial v}{\partial \nu} + \mathcal{R}, \tag{2.11}$$

where

$$\mathcal{R} = \int_\Omega 2\langle \nabla\vec{\alpha}\nabla v, \nabla v\rangle - \mathrm{div}\vec{\alpha}|\nabla v|^2 + 2\langle \vec{\alpha}, \nabla v\rangle \Delta v.$$

Here $T(P)$ is the tangent vector to $\partial\Omega$ at a point P. If we substitute $v = u$ in the above identity (2.11), then we obtain

$$\int_{\partial\Omega} \left|\frac{\partial u}{\partial T}\right|^2 \le C\left(\int_{\partial\Omega} |g|^2 + \int_{\Omega} |\nabla u|^2\right), \tag{2.12}$$

where C is depending only on $\|\vec{\alpha}\|_{C^1(\mathbb{R}^2)}$ and c_1. Since

$$\int_{\Omega} (1 + (k-1)\chi_D)|\nabla u|^2 = \int_{\partial\Omega} gu \le \left(\int_{\partial\Omega} |g|^2\right)^{1/2} \left(\int_{\partial\Omega} |u|^2\right)^{1/2}, \tag{2.13}$$

it follows from (2.12) and the Poincaré inequality on $\partial\Omega$ that

$$\int_{\Omega} |\nabla u|^2 \le C \int_{\partial\Omega} |g|^2 \tag{2.14}$$

and therefore

$$\int_{\partial\Omega} \left|\frac{\partial u}{\partial T}\right|^2 \le C \int_{\partial\Omega} |g|^2. \tag{2.15}$$

Hence (2.10) follows from (2.3), (2.14), and (2.15). This completes the proof. □

The following theorem will be used in section 6 (for the proof, see [FJR] or [V]).

THEOREM 2.2. *Let u be the solution to the Neumann problem:*

$$\begin{cases} \Delta u = 0 & \text{in } D, \\ \dfrac{\partial u}{\partial \nu} = g & \text{on } \partial D, \quad \displaystyle\int_{\partial D} u = 0. \end{cases}$$

Then there is a positive constant C depending only on the Lipschitz character (r_0, L) of D so that

$$\|(\nabla u)^*\|_{L^2(\partial D)} \le C\|g\|_{L^2(\partial D)}, \tag{2.16}$$

where

$$(\nabla u)^*(x) := \sup\{|\nabla u(y)| : |y - x| \le (2L+1)\mathrm{dist}(y, \partial D),\ y \in D\}. \tag{2.17}$$

3. Stability of disks. Recall that $\mathcal{C}[0]$ denotes the family of all disks contained in Ω_0 with radii larger than d_0. In this section we improve the logarithmic stability obtained in [KSS1] to get a Hölder stability of the disk, which we state as follows.

THEOREM 3.1. *Let g be a Neumann data with the condition (N). There exist $\alpha > 0$ and C depending only on δ_0 and d_0 such that*

$$|D_1 \Delta D_2| \le C\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|^{\alpha}_{L^\infty(\partial\Omega)} \tag{3.1}$$

for every disk $D_1, D_2 \in \mathcal{C}[0]$.

To prove (3.1), we need to study the harmonic extension property of the solution u of the Neumann problem $P[D, g]$.

LEMMA 3.2. *Let $D = B_d(a) \in \mathcal{C}[0]$. Then the general solution to $P[D, g]$ is given by*

$$\begin{cases} u^i(x) = H(x) - \dfrac{1}{2\lambda}\left(H(x) - H(a)\right), \quad x \in D, \\ u^e(x) = H(x) - \dfrac{1}{2\lambda}\left(H\left(a + \dfrac{d^2(x-a)}{|x-a|^2}\right) - H(a)\right), \quad x \in \Omega \setminus D, \end{cases} \tag{3.2}$$

where H is a harmonic function in Ω given in (2.2).

Proof. Let $u = H(x) + \mathcal{S}_D\varphi_D$ as in (2.1). Since D is a disk, $\mathcal{K}_D^*\varphi_D = 0$ on ∂D and

$$\frac{\partial u^i}{\partial \nu} = \frac{\partial H}{\partial \nu} - \frac{1}{2}\varphi = \left(1 - \frac{1}{2\lambda}\right)\frac{\partial H}{\partial \nu}$$

(see [KS1]). Note that $\int_{\partial D} \mathcal{S}_D\phi_D = \int_{\partial D} \phi_D \mathcal{S}_D 1 = 0$ because $\mathcal{S}_D 1$ is constant on ∂D. From the uniqueness of the Neumann problem in D, u^i must be as in (3.2). Now it is straightforward to check that the function u in (3.2) satisfies the transmission condition (2.9) and continuity across ∂D. This completes the proof. ∎

Using Lemma 3.2, we can derive the following lemma.

LEMMA 3.3. *Let $D = B_d(a) \in \mathcal{C}[0]$. Let u be the solution of $P[D, g]$. Then u^e extends harmonically to $\Omega \setminus \overline{B_{d-2s}(a)}$ where $s = \frac{d\delta_0}{d+2\delta_0}$; that is, there is a harmonic function $\tilde{u}^e$ in $\Omega \setminus \overline{B_{d-2s}(a)}$ so that $\tilde{u}^e = u^e$ in $\Omega \setminus \overline{D}$. Moreover, there is a positive constant C depending only on d_0 and δ_0 so that*

$$\sup_{\Omega_0 \setminus \overline{B_{d-s}(a)}} |\nabla \tilde{u}^e(x)| \le C\|g\|_{L^2(\partial\Omega)}. \tag{3.3}$$

Proof. It follows from (3.2) that

$$u^e(x) = H(x) - \frac{1}{2\lambda}\left(H\left(a + \frac{d^2(x-a)}{|x-a|^2}\right) - H(a)\right), \quad x \in \Omega \setminus D.$$

Since H is harmonic in $B_{d+2\delta_0}(a) \subset \Omega$, the first statement of Lemma 3.3 is easy. Using the previous identity and Lemma 2.1, we have

$$\sup_{\Omega_0 \setminus \overline{B_{d-s}(a)}} |\nabla \tilde{u}^e(x)| \le C \sup_{\Omega_0} |\nabla H| \le C\left(\int_\Omega |\nabla H|^2\right)^{1/2} \le C\|g\|_{L^2(\partial\Omega)}.$$

This completes the proof. ∎

PROPOSITION 3.4. *Let $D_j \in \mathcal{C}[0]$ and u_j be the solution to $P[D_j, g], j = 1, 2$. There exist constants C and α depending only on δ_0 and d_0 such that if $\|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^\infty(\partial\Omega)} \le \epsilon$, then*

$$\sup_{\Omega_0} |u_1 - u_2| \le C\epsilon^\alpha, \tag{3.4}$$

$$\sup_{\Omega_0 \setminus (D_1 \cup D_2)} |\nabla(u_1^e - u_2^e)| \le C\epsilon^\alpha, \tag{3.5}$$

$$\sup_{D_1 \cap D_2} |\nabla(u_1^i - u_2^i)| \le C\epsilon. \tag{3.6}$$

Proof. Let $D_j = B_{d_j}(a_j)$. Then $d_j > d_0$ and $\operatorname{dist}(a_j, \partial\Omega) > d_j + 2\delta_0$ because $D_j \in \mathcal{C}[0]$ $(j = 1, 2)$. By Lemma 3.3, u_j^e can be harmonically extended to $\Omega \setminus B_{d_j - 2s_j}(a_j)$, where $s_j = \frac{d_j \delta_0}{d_j + 2\delta_0}$. We also denote the extended function by u_j^e. Let v_j^e be a harmonic conjugate of u_j^e with $v_1^e = v_2^e$ on $\partial\Omega$. (Such harmonic conjugates $v_j^e (j = 1, 2)$ exist since u_j is the solution of $P[D_j, g]$ with the same Neumann data g. See [AIP].) The function $h := u_1^e + iv_1^e - (u_2^e + iv_2^e)$ is holomorphic in $\Omega \setminus \overline{B_{d_j - s_1}(a_1) \cup B_{d_j - s_2}(a_2)}$ and satisfies

$$M := \sup\{|\nabla h(x)| : x \in \Omega_0 \setminus \overline{B_{d_j - s_1}(a_1) \cup B_{d_j - s_2}(a_2)}\} \le C\|g\|_{L^2(\partial\Omega)},$$
$$|h| \le \epsilon \quad \text{on } \partial\Omega.$$

Hence M depends only on $\|g\|_{L^2(\partial\Omega)}$, δ_0, and d_0. Let ω be the solution to the Dirichlet problem:

$$\begin{cases} \Delta\omega &= 0 \quad \text{in } \Omega \setminus \overline{B_{d_1 - s_1}(a_1) \cup B_{d_2 - s_2}(a_2)}, \\ \omega &= 1 \quad \text{on } \partial\Omega, \\ \omega &= 0 \quad \text{on } \partial(B_{d_1 - s_1}(a_1) \cup B_{d_2 - s_2}(a_2)). \end{cases}$$

Then by the maximum principle, for all $x \in \Omega \setminus (D_1 \cup D_2)$,

$$\log|h(x)| \le \omega(x) \log\epsilon + (1 - \omega(x)) \log M. \tag{3.7}$$

From a standard argument, one can see that there is a positive constant α depending only on δ_0, d_0, and Ω so that

$$\inf_{\Omega \setminus (D_1 \cup D_2)} \omega \ge \alpha.$$

Therefore, we obtain

$$\sup_{\Omega \setminus (D_1 \cup D_2)} |u_1 - u_2| \le C\epsilon^\alpha.$$

The rest of the proof is the same as the proof of Proposition 5.2 of [KSS1]. This completes the proof. □

Considering Proposition 3.4, Theorem 3.1 can also be proved in the exact same way as the proof of Theorem 5.1 of [KSS1].

4. Uniform stability for the direct problem. Let Ω and Ω_0 be as before. Let D be a simply connected subdomain (not necessarily a disk) of Ω_0 with C^2-boundary. Let D_ϵ be an ϵ-perturbation of D; i.e., ∂D_ϵ is C^2 and

$$\partial D_\epsilon : P + \epsilon\omega_\epsilon(P)\nu(P), \quad P \in \partial D, \tag{4.1}$$

where $\|\omega_\epsilon\|_{C^1(\partial D)} \le 1$ and $\nu(P)$ is the outward unit normal to ∂D at P. Let u and u_ϵ be solutions of $P[D, g]$ and $P[D_\epsilon, g]$, respectively.

The main result of this section is the following uniform stability of solutions to the direct problem.

THEOREM 4.1. *Let D_ϵ be an ϵ-perturbation of the C^2-domain D. There is a positive constant C depending only on the C^2-character of D, Ω and δ_0 such that*

$$\|u - u_\epsilon\|_{L^\infty(\Omega)} < C\epsilon\|g\|_{L^2(\partial\Omega)}. \tag{4.2}$$

To prove Theorem 4.1, we need L^2-stability. The L^2-stability $\|u-u_\epsilon\|_{L^2(\Omega)} < C\epsilon$ was proven in [BF] and [BFI]. However, we will give the proof to clarify the dependency of the constant C. The following theorem states the L^2-stability.

THEOREM 4.2. *There is a positive constant C depending only on the Lipschitz characters of D, D_ϵ and δ_0 such that*

$$\|u-u_\epsilon\|_{L^2(\Omega)} < C\epsilon\|g\|_{L^2(\partial\Omega)}. \tag{4.3}$$

Indeed, the above estimate is true for general Lipschitz domains D, D_ϵ with the Hausdorff distance $\operatorname{dist}(D, D_\epsilon) < \epsilon$.

Proof. Put

$$U_\epsilon = \frac{u-u_\epsilon}{\epsilon}.$$

Let $c_\epsilon = \frac{1}{|\Omega|}\int_\Omega U_\epsilon$. We first show that

$$\|U_\epsilon - c_\epsilon\|_{L^2(\Omega)} < C\|g\|_{L^2(\partial\Omega)}. \tag{4.3$'$}$$

By integrating by parts, we have

$$\int_\Omega (1+(k-1)\chi_{D_\epsilon})\nabla U_\epsilon \nabla\xi = \frac{k-1}{\epsilon}\int_\Omega (\chi_{D_\epsilon}-\chi_D)\nabla u\nabla\xi \tag{4.4}$$

for all $\xi \in W^{1,2}(\Omega)$. Let w_ϵ be the solution to the problem:

$$\begin{cases} \nabla\cdot((1+(k-1)\chi_{D_\epsilon})\nabla w_\epsilon) = U_\epsilon - c_\epsilon \quad \text{in } \Omega, \\ \dfrac{\partial w_\epsilon}{\partial\nu} = 0 \quad \text{on } \partial\Omega, \qquad \displaystyle\int_\Omega w_\epsilon = 0. \end{cases}$$

By substituting $\eta = w_\epsilon$ in (4.4) as in [BFI], one obtains the following identity:

$$\int_\Omega |U_\epsilon - c_\epsilon|^2 = \frac{k-1}{\epsilon}\int_\Omega (\chi_{D_\epsilon}-\chi_D)\nabla u\nabla w_\epsilon.$$

Hence,

$$\int_\Omega |U_\epsilon - c_\epsilon|^2 \le \frac{|k-1|}{\epsilon}\left(\int_\Omega |\chi_{D_\epsilon}-\chi_D||\nabla u|^2\right)^{1/2}\left(\int_\Omega |\chi_{D_\epsilon}-\chi_D||\nabla w_\epsilon|^2\right)^{1/2}. \tag{4.5}$$

From (2.8) and Lemma 2.1, we obtain

$$\int_\Omega |\chi_{D_\epsilon}-\chi_D||\nabla u|^2 \le C\epsilon\|(\nabla u)^{**}\|^2_{L^2(\partial D)} \le C\epsilon\|\nabla u\|^2_{L^2(\Omega)} \le C\epsilon\|g\|^2_{L^2(\partial\Omega)}.$$

Thus, to prove (4.3$'$) it suffices to prove that

$$\int_\Omega |\chi_{D_\epsilon}-\chi_D||\nabla w_\epsilon|^2 \le C\epsilon\|U_\epsilon - c_\epsilon\|^2_{L^2(\Omega)}. \tag{4.6}$$

Let

$$V(x) := w_\epsilon(x) - \Gamma(x) + \mathcal{S}_{D_\epsilon}\eta(x), \quad x\in\Omega,$$

where

$$\Gamma(x) := \frac{1}{2\pi}\int_\Omega \log|x-y|\left(\chi_{\Omega\setminus\overline{D_\epsilon}}(y)+\frac{1}{k}\chi_{D_\epsilon}(y)\right)(U_\epsilon(y)-c_\epsilon)dy,$$

$$(\lambda I - \mathcal{K}_{D_\epsilon}^*)\eta = -\frac{\partial\Gamma}{\partial\nu} \quad \text{on } \partial D_\epsilon. \tag{4.7}$$

Then one can check that V satisfies the transmission condition (2.9) as well as $\Delta V = 0$ in $\Omega\setminus\partial D_\epsilon$. Therefore, V is the solution to the Neumann problem $P[D_\epsilon,\tilde{g}]$, where

$$\tilde{g} := \frac{\partial V}{\partial\nu} = -\frac{\partial\Gamma}{\partial\nu} + \frac{\partial}{\partial\nu}\mathcal{S}_{D_\epsilon}\eta \quad \text{on } \partial\Omega.$$

By (2.8) and Lemma 2.1, we have

$$\|(\nabla V)^{**}\|_{L^2(\partial D_\epsilon)} \le C\|\nabla V\|_{L^2(\Omega)} \le C\|\tilde{g}\|_{L^2(\partial\Omega)}. \tag{4.8}$$

From the Calderón and Zygmund estimate, we have

$$\|\Gamma\|_{W^{2,2}(\Omega)} \le C\|U_\epsilon - c_\epsilon\|_{L^2(\Omega)}. \tag{4.9}$$

It also follows from the singular integral estimate and (4.7) (see [V]) that

positive constant C depending only on the Lipschitz character of D_ϵ so that

$$\|(\nabla\mathcal{S}_{D_\epsilon}\eta)^{**}\|_{L^2(\partial D_\epsilon)} \le C\|\eta\|_{L^2(\partial D_\epsilon)} \le C\|\nabla\Gamma\|_{L^2(\partial D_\epsilon)}. \tag{4.10}$$

By (4.9), (4.10), and the trace theorem,

$$\begin{aligned}\int_{\partial\Omega}|\tilde{g}|^2 &\le C\int_{\partial\Omega}\left|\frac{\partial\Gamma}{\partial\nu}\right|^2 + \left|\frac{\partial}{\partial\nu}\mathcal{S}_{D_\epsilon}\eta\right|^2 \\ &\le C\left(\int_{\partial\Omega}|\nabla\Gamma|^2 + \int_{\partial D_\epsilon}|\nabla\Gamma|^2\right) \\ &\le C\|\Gamma\|^2_{W^{2,2}(\Omega)} \le C\|U_\epsilon - c_\epsilon\|^2_{L^2(\Omega)}.\end{aligned} \tag{4.11}$$

By (4.8) and (4.11), we obtain

$$\int_\Omega |\chi_{D_\epsilon}-\chi_D||\nabla V|^2 \le C\epsilon\|(\nabla V)^{**}\|^2_{L^2(\partial D_\epsilon)} \le C\epsilon\|\tilde{g}\|^2_{L^2(\partial\Omega)} \le C\epsilon\|U_\epsilon - c_\epsilon\|^2_{L^2(\Omega)}. \tag{4.12}$$

It then follows from (4.10) and (4.12) that

$$\begin{aligned}\int_\Omega |\chi_{D_\epsilon}-\chi_D||\nabla w_\epsilon|^2 &\le C\int_\Omega |\chi_{D_\epsilon}-\chi_D|(|\nabla V|^2 + |\nabla\Gamma|^2 + |\nabla\mathcal{S}_{D_\epsilon}\eta|^2) \\ &\le C\epsilon\|U_\epsilon - c_\epsilon\|^2_{L^2(\Omega)}.\end{aligned}$$

This proves (4.6) and (4.3′).

Now we prove (4.3). Let $\eta := \operatorname{sign}(u-u_\epsilon)$ on $\partial\Omega$ and v_η be the solution to $P[D_\epsilon, \eta - \int_{\partial\Omega}\eta]$. Then by substituting $\xi = v_\eta$ in (4.4) we obtain

$$\begin{aligned}\int_{\partial\Omega}|U_\epsilon| &= \frac{k-1}{\epsilon}\int_\Omega(\chi_{D_\epsilon}-\chi_D)\nabla u\nabla v_\eta \\ &\le C|k-1|\|(\nabla u)^{**}\|_{L^2(\partial D)}\|(\nabla v_\eta)^{**}\|_{L^2(\partial D_\epsilon)}.\end{aligned}$$

By Lemma 2.1, $\int_{\partial D_\epsilon} |(\nabla v_\eta)^{**}|^2 \le \int_{\partial D_\epsilon} |\eta - \int_{\partial\Omega} \eta|^2 \le C$ and hence we have

$$\int_{\partial\Omega} |U_\epsilon| \le C|k-1| \|\nabla u\|_{L^2(\Omega)} \le C|k-1| \|g\|_{L^2(\partial\Omega)}. \tag{4.13}$$

Let

$$\tilde{\Gamma}(x) := \frac{1}{2\pi} \int_\Omega \log|x-y| \left(\chi_{\Omega\setminus\overline{D}_\epsilon}(y) + \frac{1}{k}\chi_{D_\epsilon}(y) \right) dy$$

as in (4.6). Then, as before,

$$\tilde{V}(x) := \tilde{\Gamma}(x) + \mathcal{S}_{D_\epsilon}(\lambda I - \mathcal{K}^*_{D_\epsilon})^{-1} \left(\frac{\partial \tilde{\Gamma}}{\partial \nu}|_{\partial D_\epsilon} \right)(x)$$

satisfies $\nabla \cdot ((1+(k-1)\chi_{D_\epsilon})\nabla \tilde{V}) = 1$ in Ω and $\|\nabla \tilde{V}\|_{L^\infty(\partial\Omega)} \le C$. By substituting $\eta = \tilde{V}$ in (4.4) again, we obtain

$$\int_\Omega U_\epsilon \le \int_{\partial\Omega} |U_\epsilon| \left| \frac{\partial \tilde{V}}{\partial \nu} \right| + C\frac{1}{\epsilon} \int_\Omega |\chi_{D_\epsilon} - \chi_D| |\nabla u| |\nabla \tilde{V}|.$$

Hence (4.3) follows from the above estimate, (4.13), and (4.3′). This completes the proof. □

LEMMA 4.3. *Let* $\Omega_1 = \{X \in \Omega : \operatorname{dist}(X, \partial\Omega) < \delta_0\}$. *Then*

$$\sup_{X\in\Omega_1} |U_\epsilon(X)| + \sup_{X\in\Omega_1} |\nabla U_\epsilon(X)| \le C\|g\|_{L^2(\partial\Omega)}, \tag{4.14}$$

where the constant C *depends on* δ_0, *the Lipschitz characters of* D, D_ϵ, *and the* C^2*-character of* Ω.

Proof. Since $\Delta U_\epsilon = 0$ in Ω_1, U_ϵ and $|\nabla U_\epsilon|$ cannot assume an interior maximum in Ω_1. Suppose $X_0 \in \partial\Omega_1 \cap \Omega$. By the mean value theorem and the standard interior estimates of derivatives, we obtain

$$|U_\epsilon(X_0)| + |\nabla U_\epsilon(X_0)| \le C \left(\frac{1}{|B_{\delta_0(X_0)}|} \int_{B_{\delta_0}(X_0)} |U_\epsilon|^2 + |\nabla U_\epsilon|^2 \right)^{1/2}.$$

By Theorem 4.2 and the standard interior estimate, we obtain

$$|U_\epsilon(X_0)| + |\nabla U_\epsilon(X_0)| < C\|g\|_{L^2(\partial\Omega)}. \tag{4.15}$$

Now suppose $X_0 \in \partial\Omega$. Then we first straighten a boundary portion near X_0, and using the condition $\frac{\partial U_\epsilon}{\partial \nu} = 0$ on $\partial\Omega$ and the Schauder estimate we obtain the estimate (4.14) (see [GT, p. 126]). This completes the proof. □

By (2.1), the solutions u and u_ϵ can be expressed uniquely as

$$u = H + \mathcal{S}_D(\varphi_D) \quad \text{and} \quad u_\epsilon = H_\epsilon + \mathcal{S}_{D_\epsilon}(\varphi_{D_\epsilon}) \quad \text{in } \Omega, \tag{4.16}$$

where H, φ_D, H_ϵ, and φ_{D_ϵ} satisfy the relations (2.3) and (2.4). For notational simplicity, we will write

$$\varphi = \varphi_D,\ \mathcal{S} = \mathcal{S}_D,\ \mathcal{K}^* = \mathcal{K}^*_D,\ \varphi_\epsilon = \varphi_{D_\epsilon},\ \mathcal{S}_\epsilon = \mathcal{S}_{D_\epsilon},\ \mathcal{K}^*_\epsilon = \mathcal{K}^*_{D_\epsilon}.$$

We first prove the stability of the harmonic part of the solution in Lemma 4.4.

LEMMA 4.4. *We have that*

$$\|H - H_\epsilon\|_{L^\infty(\Omega)} + \|\nabla(H - H_\epsilon)\|_{L^\infty(\Omega)} < C\epsilon\|g\|_{L^2(\partial\Omega)}, \tag{4.17}$$

where the constant C depends on δ_0, the Lipschitz characters of D, D_ϵ, and the C^2-character of Ω.

Proof. From (2.3), we have

$$\frac{H - H_\epsilon}{\epsilon} = \mathcal{D}_\Omega(U_\epsilon|_{\partial\Omega}).$$

Since $\partial\Omega \in C^2$, $|\frac{\langle P-Q,\nu(Q)\rangle}{|P-Q|^2}| \le C$ for $P, Q \in \partial\Omega$ where C depends on the C^2-character of $\partial\Omega$. Hence for $P \in \partial\Omega$,

$$\begin{aligned}\left|\frac{H(P) - H_\epsilon(P)}{\epsilon}\right| &\le \int_{\partial\Omega}\left|\frac{\partial}{\partial\nu}\Gamma(P-Q)\right||U_\epsilon(Q)|d\sigma(Q)\\ &\le C\|U_\epsilon\|_{L^1(\partial\Omega)} \le C\|g\|_{L^2(\partial\Omega)}\end{aligned}$$

by Lemma 4.3. Thus by the maximum principle we have

$$\left\|\frac{H - H_\epsilon}{\epsilon}\right\|_{L^\infty(\Omega)} \le C\|g\|_{L^2(\partial\Omega)}.$$

By the relation (4.16) and Lemma 4.3, we have

$$\left\|\frac{\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon}{\epsilon}\right\|_{L^\infty(\partial\Omega)} \le \left\|\frac{H - H_\epsilon}{\epsilon}\right\|_{L^\infty(\partial\Omega)} + \|U_\epsilon\|_{L^\infty(\partial\Omega)} \le C\|g\|_{L^2(\partial\Omega)}.$$

Since $\Delta(\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon) = 0$ in $\mathbb{R}^2 \setminus (\partial D \cup \partial D_\epsilon)$ and $|\mathcal{S}\varphi(X)| = O(|X|^{-1})$ for $|X|$ large (recall $\int_{\partial D}\varphi = 0$),

$$\left\|\frac{\mathcal{S}\varphi - \mathcal{S}\varphi_\epsilon}{\epsilon}\right\|_{L^\infty(\mathbb{R}^2\setminus\Omega)} = \left\|\frac{\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon}{\epsilon}\right\|_{L^\infty(\partial\Omega)} \le C\|g\|_{L^2(\partial\Omega)}. \tag{4.18}$$

Then, for $P \in \partial\Omega$,

$$\begin{aligned}&\left|\nabla\left(\frac{\mathcal{S}\varphi - S_\epsilon\varphi_\epsilon}{\epsilon}\right)(P)\right|\\ &\le C\left\|\frac{\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon}{\epsilon}\right\|_{L^1(B_{\delta_0}(P))} \qquad \text{(by the interior estimate)}\\ &\le C\left\|\frac{\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon}{\epsilon}\right\|_{L^1(B_{\delta_0}(P)\cap\Omega)} + C\|g\|_{L^2(\partial\Omega)} \qquad \text{(by (4.18))}\\ &\le C\left\|\frac{H - H_\epsilon}{\epsilon}\right\|_{L^1(B_{\delta_0}(P)\cap\Omega)} + \|U_\epsilon\|_{L^1(B_{\delta_0}(P)\cap\Omega)} + C\|g\|_{L^2(\partial\Omega)} \qquad \text{(by (4.16))}\\ &\le C\|g\|_{L^2(\partial\Omega)}.\end{aligned}$$

From Lemma 4.3 and (4.16), we obtain

$$\begin{aligned}\left\|\nabla\left(\frac{H - H_\epsilon}{\epsilon}\right)\right\|_{L^\infty(\Omega)} &\le \left\|\nabla\left(\frac{H - H_\epsilon}{\epsilon}\right)\right\|_{L^\infty(\partial\Omega)}\\ &\le \|\nabla U_\epsilon\|_{L^\infty(\partial\Omega)} + \left\|\nabla\left(\frac{\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon}{\epsilon}\right)\right\|_{L^\infty(\partial\Omega)}\\ &\le C\|g\|_{L^2(\partial\Omega)}.\end{aligned}$$

This completes the proof. □

We now prove the uniform stability of the single layer potential which, combined with Lemma 4.4, proves Theorem 4.1.

We first compare $\mathcal{K}_\epsilon^*$ with $\mathcal{K}^*$ in the uniform norm. Let Φ_ϵ be the diffeomorphism from ∂D onto ∂D_ϵ given by $\Phi_\epsilon(P) = P + \epsilon\omega_\epsilon(P)\nu(P)$.

LEMMA 4.5. *There exists C depending only on the C^2-character of ∂D such that for any function $f \in L^2(\partial D_\epsilon)$,*

$$\|(\mathcal{K}_\epsilon^* f)\circ\Phi_\epsilon - \mathcal{K}^*(f\circ\Phi_\epsilon)\|_{L^2(\partial D)} \le C\epsilon\|f\|_{L^2(\partial D_\epsilon)}. \tag{4.19}$$

Proof. Fix $P \in \partial D$. Let

$$k(P,Q) = \frac{\langle P-Q, \nu(P)\rangle}{|P-Q|^2} \quad \text{and} \quad k_\epsilon(P,Q) = \frac{\langle \Phi_\epsilon(P)-\Phi_\epsilon(Q), \nu(\Phi_\epsilon(P))\rangle}{|\Phi_\epsilon(P)-\Phi_\epsilon(Q)|^2},$$

which are integral kernels for $\mathcal{K}$ and $\mathcal{K}_\epsilon$. Then

$$\begin{aligned}&(\mathcal{K}_\epsilon^* f)\circ\Phi_\epsilon(P) - \mathcal{K}^*(f\circ\Phi_\epsilon)(P)\\ &\quad = \int_{\partial D}[k_\epsilon(P,Q)j_\epsilon(Q) - k(P,Q)]f\circ\Phi_\epsilon(Q)d\sigma(Q),\end{aligned}$$

where j_ϵ is the Jacobian of Φ_ϵ. To estimate $k_\epsilon(P,Q) - k(P,Q)$, we observe that

$$\nu(\Phi_\epsilon(P)) = (1+O(\epsilon))\nu(P) + O(\epsilon)T(P), \tag{4.20}$$

where $T(P)$ is the unit tangent to ∂D at P. Note that above $O(\epsilon)$ depends only on the C^2-norm of ∂D. Since $\partial D \in C^2$, we have

$$k_\epsilon(P,Q) - k(P,Q) = O(\epsilon) + O(\epsilon)\frac{\langle P-Q, T(P)\rangle}{|P-Q|^2}$$

and hence

$$\begin{aligned}&|(\mathcal{K}_\epsilon^* f)\circ\Phi_\epsilon(P) - \mathcal{K}^*(f\circ\Phi_\epsilon)(P)|\\ &\quad \le O(\epsilon)\int_{\partial D}|f\circ\Phi_\epsilon(Q)|d\sigma(Q) + O(\epsilon)|\mathcal{T}(f\circ\Phi_\epsilon)(P)|,\end{aligned}$$

where

$$\mathcal{T}f\circ\Phi_\epsilon(P) = \int_{\partial D}\frac{\langle P-Q, T(P)\rangle}{|P-Q|^2}f\circ\Phi_\epsilon(Q)d\sigma(Q).$$

It is now standard to show that

$$\|\mathcal{T}f\|_{L^2(\partial D)} \le C\|f\|_{L^2(\partial D)}.$$

This completes the proof. □

LEMMA 4.6. *We have that*

$$\|\varphi_\epsilon\circ\Phi_\epsilon - \varphi\|_{L^2(\partial D)} \le C\epsilon\|g\|_{L^2(\partial\Omega)}, \tag{4.21}$$

where the constant C depends on C^2-characters of $\partial D, \partial\Omega$, and δ_0.

Proof. Let $\lambda = \frac{k+1}{2(k-1)}$. Since $(\lambda I - \mathcal{K}^*)$ is invertible on $L^2(\partial D)$ (see [F]), we have from Lemmas 4.4 and 4.5 that

$$\begin{aligned}
&\|\varphi_\epsilon \circ \Phi_\epsilon - \varphi\|_{L^2(\partial D)} \\
&\quad \le C\|(\lambda I - \mathcal{K}^*)(\varphi_\epsilon \circ \Phi_\epsilon - \varphi)\|_{L^2(\partial D)} \\
&\quad \le C\|((\lambda I - \mathcal{K}_\epsilon^*)\varphi_\epsilon) \circ \Phi_\epsilon - (\lambda I - \mathcal{K}^*)\varphi\|_{L^2(\partial D)} \\
&\qquad\quad + C\|(\mathcal{K}_\epsilon^*\varphi_\epsilon) \circ \Phi_\epsilon - \mathcal{K}^*(\varphi_\epsilon \circ \Phi_\epsilon)\|_{L^2(\partial D)} \\
&\quad \le C\left\|\frac{\partial H_\epsilon}{\partial \nu} \circ \Phi_\epsilon - \frac{\partial H}{\partial \nu}\right\|_{L^2(\partial D)} + C\epsilon\|\varphi_\epsilon\|_{L^2(\partial D_\epsilon)} \\
&\quad \le C\left\|\frac{\partial H_\epsilon}{\partial \nu} \circ \Phi_\epsilon - \frac{\partial H}{\partial \nu} \circ \Phi_\epsilon\right\|_{L^2(\partial D)} + C\left\|\frac{\partial H}{\partial \nu} \circ \Phi_\epsilon - \frac{\partial H}{\partial \nu}\right\|_{L^2(\partial D)} \\
&\qquad\quad + C\epsilon\left\|(\lambda I - \mathcal{K}_\epsilon^*)^{-1}\frac{\partial H_\epsilon}{\partial \nu}\right\|_{L^2(\partial D_\epsilon)} \\
&\quad \le C\epsilon(\|H\|_{W^{2,\infty}(\Omega_0)} + \|H\|_{W^{2,\infty}(\Omega_0)}) \\
&\quad \le C\epsilon\left(\|H\|_{L^2(\Omega)} + \|H_\epsilon\|_{L^2(\Omega)}\right) \\
&\quad \le C\epsilon\|g\|_{L^2(\partial\Omega)}.
\end{aligned}$$

The last inequality follows from Lemma 2.1. This completes the proof. ∎

LEMMA 4.7. *We have that*

$$\|\mathcal{S}\varphi - \mathcal{S}_\epsilon\varphi_\epsilon\|_{L^\infty(\Omega)} \le C\epsilon\|g\|_{L^2(\partial\Omega)},$$

where the constant C depends on C^2-characters of $\partial D, \partial\Omega$, and δ_0.

Proof. Since $\mathcal{S}\varphi(X) = \mathcal{S}_\epsilon\varphi_\epsilon(X) = O(|X|^{-1})$ as $|X| \to \infty$, by the maximum principle, there is a $P \in \partial D \cup \partial D_\epsilon$ such that

$$|\mathcal{S}\varphi(P) - \mathcal{S}_\epsilon\varphi_\epsilon(P)| = \sup_{X\in\Omega} |\mathcal{S}\varphi(X) - \mathcal{S}_\epsilon\varphi_\epsilon(X)|.$$

Suppose that $P \in \partial D$. (The case when $P \in \partial D_\epsilon$ can be treated in the exact same way by interchanging the role of ∂D and ∂D_ϵ.) Then

$$\begin{aligned}
&\mathcal{S}\varphi(P) - \mathcal{S}_\epsilon\varphi_\epsilon(P) \\
&= \frac{1}{2\pi}\int_{\partial D} [\log|P - Q|\varphi(Q) - \log|P - \Phi_\epsilon(Q)|]\, j_\epsilon(Q)\varphi_\epsilon \circ \Phi_\epsilon(Q)d\sigma(Q) \\
&= \frac{1}{2\pi}\int_{\partial D} [\log|P - Q| - \log|P - \Phi_\epsilon(Q)|]\, \varphi(Q)d\sigma(Q) \\
&\quad + \frac{1}{2\pi}\int_{\partial D} \log|P - \Phi_\epsilon(Q)|\, [1 - j_\epsilon(Q)]\, \varphi_\epsilon \circ \Phi_\epsilon(Q)d\sigma(Q) \\
&\quad + \frac{1}{2\pi}\int_{\partial D} \log|P - \Phi_\epsilon(Q)|\, [\varphi(Q) - \varphi_\epsilon \circ \Phi_\epsilon(Q)]\, d\sigma(Q) \\
&:= I_1 + I_2 + I_3.
\end{aligned}$$

It follows from Lemma 4.6 that

$$|I_3| \le C\epsilon\|g\|_{L^2(\partial\Omega)}.$$

It is easy to see that

$$|I_2| \le C\epsilon\|g\|_{L^2(\partial\Omega)}.$$

Since

$$\left| \int_{\partial D} \log|P-Q| - \log|P-\Phi_\epsilon(Q)| d\sigma(Q) \right| \le C\epsilon,$$

it follows from the standard argument of singular integral that

$$|I_1| \le C\epsilon \|\varphi\|_{C^{1/2}(\partial D)} \le C\epsilon \|g\|_{L^2(\partial\Omega)}.$$

The last inequality follows from Lemma 2.1 and the invertibility of $\lambda I - \mathcal{K}^*$ on $C^\alpha(\partial D)$, $0 < \alpha < 1$. This completes the proof. □

5. Perturbation of disk-error estimates. In this section, we apply the results from sections 3 and 4 to the class of perturbation of disks. From Theorem 4.1, we have the following theorem.

THEOREM 5.1. *Let $D \in \mathcal{C}[\epsilon]$ and D be an ϵ-perturbation of a disk B. There is a constant C depending on δ_0 and d_0, not on D or B, such that if u_D and u_B are solutions to $P[D,g]$ and $P[B,g]$, respectively, then*

$$\|u_D - u_B\|_{L^\infty(\Omega)} \le C\epsilon. \tag{5.1}$$

Theorems 5.1 and 4.1 give the following stability for $\mathcal{C}[\epsilon]$.

THEOREM 5.2. *Suppose that $D_1, D_2 \in \mathcal{C}[\epsilon]$. Let g satisfy the condition (N). Then there is a positive constant C independent of D_j so that*

$$|D_1 \Delta D_2| \le C\left(\epsilon + \|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^\infty(\partial\Omega)}\right)^\alpha. \tag{5.2}$$

Proof. By the definition, there are two balls B_1 and B_2 such that D_j is an ϵ-perturbation of B_j. Let v_j and u_j be the solutions of the Neumann problems $P[B_j, g]$ and $P[D_j, g]$, respectively. By Theorem 5.1, we have

$$\sup_{x\in\Omega} |v_j(x) - u_j(x)| \le C\epsilon.$$

Hence

$$\|v_1 - v_2\|_{L^\infty(\partial\Omega)} \le C\left(\epsilon + \|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^\infty(\partial\Omega)}\right).$$

It then follows from Theorem 3.1 that

$$d(B_1, B_2) \le C\left(\epsilon + \|\Lambda_{D_1}(g) - \Lambda_{D_2}(g)\|_{L^\infty(\partial\Omega)}\right)^\alpha$$

and hence (5.2) follows. This completes the proof. □

6. Approximate identification (without restriction on g). In this section, we consider the uniqueness question without imposing any restriction on the Neumann data g. Let g be any given nonzero function on $\partial\Omega$. Let $D_0 \in \mathcal{C}[\epsilon]$ and let $\Lambda_{D_0}(g) = f$. We obtain that if $D \in \mathcal{C}[\epsilon]$ satisfies $\Lambda_D(g) = f$, then D must be close to D_0. Indeed, we obtain the following result.

THEOREM 6.1. *There is a positive constant C such that if $D \in \mathcal{C}[\epsilon]$ and $\Lambda_D(g) = f$ on $\partial\Omega$, then*

$$|D \Delta D_0| \le C\epsilon.$$

Remark. Theorem 6.1 together with the result in [AIP] gives the global uniqueness within $\mathcal{C}[\epsilon]$ for sufficiently small ϵ if g satisfies the following condition: $\partial\Omega$ is the union of two disjoint arcs Γ_1 and Γ_2, and

$$g \geq 0 \quad \text{on } \Gamma_1 \qquad \text{and} \qquad g \leq 0 \quad \text{on } \Gamma_2.$$

Here $\partial\Omega$ is smooth.

Without loss of generality, we assume D_0 is an ϵ-perturbation of a disk $B_0 = B_{r_0}(0)$ and D is an ϵ-perturbation of a disk $B = B_r(a)$. For notational simplicity, we will write

$$\varphi_0 = \varphi_{D_0},\ \varphi = \varphi_D;\ \mathcal{S}_0 = \mathcal{S}_{D_0},\ \mathcal{S} = \mathcal{S}_D;\ \mathcal{K}_0^* = \mathcal{K}_{D_0}^*,\ \mathcal{K}^* = \mathcal{K}_D^*.$$

By (2.2), the solution $u = u_D$ can be expressed uniquely as

$$u = H + \mathcal{S}\varphi \quad \text{in } \Omega, \tag{6.1}$$

where H and φ are given in the formulas (2.3) and (2.4); that is,

$$\varphi = (\lambda I - \mathcal{K}^*)^{-1}\left(\frac{\partial H}{\partial \nu}\right) \qquad \text{on } \partial D. \tag{6.2}$$

From (6.2), $\int_{\partial D} \varphi = 0 = \int_{\partial D} \mathcal{K}^*\varphi$.

LEMMA 6.2. *We have that*

$$\mathcal{S}\varphi = -\frac{1}{2\lambda}H + w + c \quad \text{in } D, \tag{6.3}$$

where c is a constant and w is the solution to the Neumann problem:

$$\begin{cases} \Delta w = 0 & \text{in } D, \\ \dfrac{\partial w}{\partial \nu}|_{\partial D} = \left(1 - \dfrac{1}{2\lambda}\right)\mathcal{K}^*\varphi, \quad \displaystyle\int_{\partial D} w = 0. \end{cases} \tag{6.4}$$

Proof. Since

$$\begin{aligned} \frac{\partial}{\partial \nu^-}\mathcal{S}\varphi|_{\partial D} &= \left(-\frac{1}{2}I + \mathcal{K}^*\right)\varphi \\ &= -\frac{1}{2\lambda}(\lambda I - \mathcal{K}^*)\varphi + \left(1 - \frac{1}{2\lambda}\right)\mathcal{K}^*\varphi \\ &= -\frac{1}{2\lambda}\frac{\partial H}{\partial \nu}\Big|\partial D + \left(1 - \frac{1}{2\lambda}\right)\mathcal{K}^*\varphi, \end{aligned} \tag{6.5}$$

(6.3) follows from the uniqueness of the Neumann problem. □

Observe that if D is a disk, then $\mathcal{K}^*\varphi = 0$ and hence $w = 0$ in D (see [KS1]).

LEMMA 6.3. *There exists C depending only on δ_0 and d_0 such that*

$$\|\mathcal{K}^*\varphi\|_{L^2(\partial D)} \leq C\epsilon\|\varphi\|_{L^2(\partial D)}.$$

Proof. Let φ_B be the density function in (6.2) with D replaced by B. Then $\mathcal{K}_B^*\varphi_B = 0$. Therefore, Lemma 6.3 follows from Lemma 4.5. □

Set

$$U = u_0 - u = \mathcal{S}_0\varphi_0 - \mathcal{S}\varphi. \tag{6.6}$$

It follows from the basic regularity theory that $u \in C^1(\overline{D}) \cap C^1(\Omega \setminus D)$ and $u_0 \in C^1(\overline{D_0}) \cap C^1(\Omega \setminus D_0)$. We will denote by $\frac{\partial U^\#}{\partial \nu}$ the outer normal derivative of U on the boundary from the region $(D \setminus \overline{D_0}) \cup (D_0 \setminus \overline{D})$.

LEMMA 6.4. *Assume that $\Omega \setminus \overline{D \cup D_0}$ is connected. Then*

(1) $U = 0$ *in* $\Omega \setminus \overline{D \cup D_0}$,

(2) $$\frac{\partial U^\#}{\partial \nu} = -\varphi \quad \text{on } \partial D \setminus \overline{D_0} \quad \text{and} \quad \frac{\partial U^\#}{\partial \nu} = -\varphi_0 \quad \text{on } \partial D_0 \setminus \overline{D},$$

(3) $$\frac{\partial U^\#}{\partial \nu} = \varphi + \frac{\partial}{\partial \nu}(w_0 - w) \quad \text{on } \partial D \cap D_0,$$
$$\frac{\partial U^\#}{\partial \nu} = \varphi_0 + \frac{\partial}{\partial \nu}(w_0 - w) \quad \text{on } \partial D_0 \cap D.$$

Here w_0 is the function in Lemma 6.2 for D_0; i.e., $\mathcal{S}_0\varphi_0 = -\frac{1}{2\lambda}H + w_0 + c_0$ in D_0.

Proof. Since $U = \frac{\partial U}{\partial \nu} = 0$ on $\partial\Omega$, (1) follows from the unique continuation of the harmonic function. From (1) and the jump relation (2.1), for $x \in \partial D_0 \setminus \overline{D}$,

$$\begin{aligned}\frac{\partial U^\#}{\partial \nu}(x) &= \frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u^e}{\partial \nu}(x) = \frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u^e}{\partial \nu}(x) - \left(\frac{\partial u_0^e}{\partial \nu}(x) - \frac{\partial u^e}{\partial \nu}(x)\right) \\ &= \frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u_0^e}{\partial \nu}(x) = -\varphi_0(x),\end{aligned}$$

which proves the second part of (2). The first part of (2) can be proved in the same way. From Lemma 6.2 and the trace formula (2.1), we have for $x \in \partial D \cap D_0$,

$$\begin{aligned}\frac{\partial U^\#}{\partial \nu}(x) &= \frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u^e}{\partial \nu}(x) \\ &= \frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u^e}{\partial \nu}(x) - \left(\frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u^i}{\partial \nu}(x)\right) + \left(\frac{\partial u_0^i}{\partial \nu}(x) - \frac{\partial u^i}{\partial \nu}(x)\right) \\ &= \varphi(x) + \frac{\partial}{\partial \nu}(w_0 - w)(x).\end{aligned}$$

This completes the proof. ∎

LEMMA 6.5. *Assume that $\Omega \setminus \overline{D \cup D_0}$ is connected. Then there exists a positive constant C depending only on the radii of B_0 and B so that*

$$\begin{aligned}&\frac{1}{2}\left(\|\varphi_0\|^2_{L^2(\partial D_0 \setminus E)} + \|\varphi\|^2_{L^2(\partial D \setminus E)}\right) - C\epsilon^2\left(\|\varphi_0\|^2_{L^2(\partial D_0)} + \|\varphi\|^2_{L^2(\partial D)}\right) \\ &\le \int_{\partial(D \setminus \overline{D_0})} \left|\frac{\partial U^\#}{\partial \nu}\right|^2 + \int_{\partial(D_0 \setminus \overline{D})} \left|\frac{\partial U^\#}{\partial \nu}\right|^2 \\ &\le C\left(\|\varphi_0\|^2_{L^2(\partial D_0)} + \|\varphi\|^2_{L^2(\partial D)}\right),\end{aligned} \tag{6.7}$$

where $E = \partial D_0 \cap \partial D$.

Proof. From Lemma 6.4, we have

$$\int_{\partial D_0\setminus\overline{D}}\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2=\int_{\partial D_0\setminus\overline{D}}|\varphi_0|^2.$$

We also obtain

$$\int_{\partial D\cap D_0}\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2\geq\frac{1}{2}\int_{\partial D\cap D_0}|\varphi|^2-\int_{\partial D\cap D_0}\left|\frac{\partial w_0}{\partial\nu}-\frac{\partial w}{\partial\nu}\right|^2.$$

Thus by Theorem 2.2 and Lemmas 6.2 and 6.3, we have

$$\begin{aligned}\int_{\partial D\cap D_0}\left|\frac{\partial w_0}{\partial\nu}-\frac{\partial w}{\partial\nu}\right|^2&\leq 2\int_{\partial D_0}|(\nabla w_0)^*|^2+2\int_{\partial D}|(\nabla w)^*|^2\\&\leq C\left(\int_{\partial D_0}\left|\frac{\partial w_0}{\partial\nu}\right|^2+\int_{\partial D}\left|\frac{\partial w}{\partial\nu}\right|^2\right)\\&=C\left(\int_{\partial D_0}|\mathcal{K}_0^*\varphi_0|^2+\int_{\partial D}|\mathcal{K}^*\varphi|^2\right)\\&\leq C\epsilon^2\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right).\end{aligned}$$

The other inequalities can be obtained in the same way. Thus we have the first inequality in (6.7).

The second inequality in (6.7) follows from the singular integral estimates

$$\|(\nabla\mathcal{S}\varphi)^{**}\|_{L^2(\partial D)}\leq C\|\varphi\|_{L^2(\partial D)}\quad\text{and}\quad\|(\nabla\mathcal{S}_0\varphi_0)^{**}\|_{L^2(\partial D)}\leq C\|\varphi_0\|_{L^2(\partial D)},$$

where C is a positive constant depending only on the Lipschitz characters of D and D_0. This completes the proof. □

We are now ready to prove Theorem 6.1.

Proof of Theorem 6.1. If $\Lambda_{D_0}(g)=\Lambda_D(g)$ on $\partial\Omega$, it is known that $\overline{D}\cap\overline{D_0}\neq\emptyset$ and one domain cannot be contained in the other (see [FI].) Hence it must be

$$\partial D\cap\partial D_0\neq\emptyset.$$

Assume first that $\Omega\setminus\overline{D\cup D_0}$ is connected. Let M be a fixed number to be chosen later. Take a function $\eta\in C^2(\mathbb{R}^2)$ so that $\eta=1$ in $\mathbb{R}^2\setminus B_{r+2M\epsilon}(a)$, $\eta=0$ in $B_{r+\epsilon}(a)$, and $\|\nabla\eta\|_{L^\infty}\leq\frac{1}{M\epsilon}$. Let $\vec{\alpha}(x)=x\eta(x)$. If we apply the Rellich identity as in (2.11) with $U=u_0-u$ over the region $D_0\setminus\overline{D}$, then

(6.8)
$$\int_{\partial(D_0\setminus\overline{D})}\langle\vec{\alpha},\nu\rangle\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2=\int_{\partial(D_0\setminus\overline{D})}\langle\vec{\alpha},\nu\rangle\left|\frac{\partial U}{\partial T}\right|^2-2\int_{\partial(D_0\setminus\overline{D})}\langle\vec{\alpha},T\rangle\frac{\partial U}{\partial T}\frac{\partial U^{\#}}{\partial\nu}+\mathcal{R},$$

where

$$\mathcal{R}=\int_{D_0\setminus\overline{D}}2\langle\nabla\vec{\alpha}\nabla U,\nabla U\rangle-\mathrm{div}\vec{\alpha}|\nabla U|^2.$$

Here $\frac{\partial U}{\partial T}$ denotes the tangential derivative on the boundary. Since $U=0$ in $\Omega\setminus\overline{D_0\cup D}$, $\frac{\partial U}{\partial T}=0$ along $\partial D_0\setminus\overline{D}$ and $\partial D\setminus\overline{D_0}$. And $\vec{\alpha}=0$ on $\partial D\cap\overline{D_0}$. Therefore, the first and

second terms of the right side of the equality in (6.8) vanish. Moreover, we have

$$
\begin{aligned}
|\mathcal{R}| &\le C\|\nabla\vec{\alpha}\|_{L^\infty(\mathbb{R}^2)}\int_{D_0\setminus\overline{D}}|\nabla U|^2 \\
&\le C\left(\frac{1}{M\epsilon}+1\right)\int_{\partial(D_0\setminus\overline{D})}\frac{\partial U^\#}{\partial\nu}U \\
&\le C\left(\frac{1}{M\epsilon}+1\right)\left(\int_{\partial(D_0\setminus\overline{D})}\left|\frac{\partial U^\#}{\partial\nu}\right|^2\right)^{1/2}\left(\int_{\partial(D_0\setminus\overline{D})}|U|^2\right)^{1/2}.
\end{aligned}
\tag{6.9}
$$

From Lemma 6.5, we obtain

$$
|\mathcal{R}| \le C\left(\frac{1}{M\epsilon}+1\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right)^{1/2}\left(\int_{\partial(D_0\setminus\overline{D})}|U|^2\right)^{1/2}. \tag{6.10}
$$

It follows from Lemma 6.2 that

$$
\begin{aligned}
\int_{\partial(D_0\setminus\overline{D})}|U|^2 &= \int_{\overline{D_0}\cap\partial D}|U|^2 \\
&= \int_{\overline{D_0}\cap\partial D}|w_0-c_0-(w-c)|^2 \\
&\le C(\|w\|^2_{L^\infty(D)}+\|w\|^2_{L^\infty(D)}+|c-c_0|^2).
\end{aligned}
$$

We now show that

$$
\|w\|^2_{L^\infty(D)}+\|w\|^2_{L^\infty(D)}+|c-c_0|^2 \le C\epsilon^2\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right). \tag{6.11}
$$

There exists $\psi\in L^2(\partial D)$ with $\int_{\partial D}\psi=0$ so that $(-\frac{1}{2}I+\mathcal{K}^*)\psi=(1-\frac{1}{2\lambda})\mathcal{K}^*\varphi$ on ∂D (see [F]). Thus

$$
\begin{aligned}
&w(x)=\frac{1}{2\pi}\int_{\partial D}\log|x-y|\psi(y)d\sigma(y)\quad\text{for } x\in D,\\
&\|\psi\|_{L^2(\partial D)}\le C\left\|\left(1-\frac{1}{2\lambda}\right)\mathcal{K}^*\varphi\right\|_{L^2(\partial D)}\le C\epsilon\|\varphi\|_{L^2(\partial D)}.
\end{aligned}
$$

From the Schwartz inequality, we obtain

$$
\|w\|_{L^\infty(D)}\le\left(\int_{\partial D}|\log|x-y||^2\right)^{1/2}\left(\int_{\partial D}|\psi|^2\right)^{1/2}\le C\epsilon\|\varphi\|_{L^2(\partial D)}.
$$

Suppose that $\partial(\overline{D_0\cup D})\cap\partial(D_0\cap D)\neq\emptyset$. At a point $P\in\partial(\overline{D_0\cup D})\cap\partial(D_0\cap D)$, $U(P)=0=\mathcal{S}_0\varphi_0(P)-\mathcal{S}\varphi(P)$. It follows from the above estimate that

$$
\begin{aligned}
|c_0-c| &= |w_0(P)-w(P)| \\
&\le \|w_0\|_{L^\infty(D_0)}+\|w\|_{L^\infty(D)} \\
&\le C\epsilon\left(\|\varphi_0\|_{L^2(\partial D_0)}+\|\varphi\|_{L^2(\partial D)}\right).
\end{aligned}
$$

This proves the estimate (6.11).

It then follows from (6.8)–(6.11) that

$$\int_{\partial(D_0\setminus\overline{D})}\langle\vec{\alpha},\nu\rangle\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2\le C\left(\frac{1}{M}+\epsilon\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right). \tag{6.12}$$

As before, we apply the Rellich identity (6.8) with the vector field $\vec{\beta}(x)=(\eta(x)-1)(x-a)$ over the domain $D_0\setminus\overline{D}$ and obtain

$$\int_{\partial(D_0\setminus\overline{D})}\langle\vec{\beta},\nu\rangle\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2\le C\left(\frac{1}{M}+\epsilon\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right)+|\mathcal{J}|, \tag{6.13}$$

where

$$\mathcal{J}=\int_{\partial(D\setminus\overline{D_0})}\langle\vec{\beta},\nu\rangle\left|\frac{\partial U}{\partial T}\right|^2-2\int_{\partial(D\setminus\overline{D_0})}\langle\vec{\beta},T\rangle\frac{\partial U}{\partial T}\frac{\partial U^{\#}}{\partial\nu}.$$

Since $\frac{\partial U}{\partial T}=0$ on $\partial D\setminus\overline{D_0}$,

$$|\mathcal{J}|\le C\int_{\partial D_0\cap\overline{D}}\left|\frac{\partial U}{\partial T}\right|^2+\left|\frac{\partial U}{\partial T}\right|\left|\frac{\partial U^{\#}}{\partial\nu}\right|. \tag{6.14}$$

Note that

$$\frac{\partial U}{\partial T}=\frac{\partial}{\partial T}(w_0-w)\quad\text{ on }\partial D_0\cap D.$$

By Lemmas 6.3 and 6.5 and (6.14), we obtain

$$|\mathcal{J}|\le C\epsilon\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right).$$

It then follows from (6.13) that

$$\int_{\partial(D_0\setminus\overline{D})}\langle\vec{\beta},\nu\rangle\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2\le C\left(\frac{1}{M}+\epsilon\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right). \tag{6.15}$$

Let $\gamma_0^\epsilon=(\partial D_0\setminus\overline{D})\cap B_{r+2M\epsilon}(a)$. Observe that $\langle\vec{\alpha},\nu\rangle\ge r_0-\epsilon$ on $(\partial D_0\setminus\overline{D})\setminus\gamma_0^\epsilon$ and $\langle\vec{\beta},\nu\rangle\ge r-\epsilon$ on $\partial D\cap\overline{D_0}$. It follows from (6.12) and (6.15) that

$$\int_{\partial(D_0\setminus\overline{D})}\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2-C\int_{\gamma_0^\epsilon}\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2\le C\left(\frac{1}{M}+\epsilon\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right). \tag{6.16}$$

If we put $\gamma^\epsilon=(\partial D\setminus\overline{D_0})\cap B_{r+2M\epsilon}(0)$, we have in the same way that

$$\int_{\partial(D\setminus\overline{D_0})}\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2-C\int_{\gamma^\epsilon}\left|\frac{\partial U^{\#}}{\partial\nu}\right|^2\le C\left(\frac{1}{M}+\epsilon\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right). \tag{6.17}$$

In (6.16) and (6.17), the constant C depends only on the radii of B_0 and B. Recall that $\frac{\partial U^{\#}}{\partial\nu}=\varphi$ on γ^ϵ and $\frac{\partial U^{\#}}{\partial\nu}=\varphi_0$ on γ_0^ϵ (see Lemma 6.4). From (6.7), (6.16), and (6.17), we obtain that

$$\begin{aligned}\|\varphi_0\|^2_{L^2(\partial D_0\setminus E)}+\|\varphi\|^2_{L^2(\partial D\setminus E)}\le{}&\left(C\epsilon+\frac{C}{M}\right)\left(\|\varphi_0\|^2_{L^2(\partial D_0)}+\|\varphi\|^2_{L^2(\partial D)}\right)\\&+C\left(\|\varphi_0\|^2_{L^2(\gamma_0^\epsilon)}+\|\varphi\|^2_{L^2(\gamma^\epsilon)}\right).\end{aligned} \tag{6.18}$$

Now let us choose M so that $\frac{C}{M} < \frac{1}{4}$. It then follows from (6.18) that

$$\|\varphi_0\|^2_{L^2(\partial D_0)} + \|\varphi\|^2_{L^2(\partial D)} \le C\left(\|\varphi_0\|^2_{L^2(\gamma_0^\epsilon \cup E)} + \|\varphi\|^2_{L^2(\gamma^\epsilon \cup E)}\right) \tag{6.19}$$

provided that ϵ is small enough so that $C\epsilon \le \frac{1}{4}$. By (6.2) and Lemma 6.3, we have

$$\begin{aligned}|\lambda|\|\varphi\|_{L^2(\gamma^\epsilon \cup E)} &\le \|\mathcal{K}^*\varphi\|_{L^2(\gamma^\epsilon \cup E)} + \|(\lambda I - \mathcal{K}^*)\varphi\|_{L^2(\gamma^\epsilon \cup E)} \\ &\le C\epsilon\|\varphi\|_{L^2(\partial D)} + \left\|\frac{\partial H}{\partial \nu}\right\|_{L^2(\gamma^\epsilon \cup E)}.\end{aligned}$$

Therefore we have from (6.2) and (6.19) that

$$\left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\partial D_0)} + \left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\partial D)} \le C\left(\left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\gamma_0^\epsilon \cup E)} + \left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\gamma^\epsilon \cup E)}\right). \tag{6.20}$$

By the interior estimate of harmonic functions, one can easily see that

$$\left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\gamma^\epsilon \cup E)} + \left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\gamma_0^\epsilon \cup E)} \le C(l(\gamma_0^\epsilon \cup E) + l(\gamma^\epsilon \cup E)) \int_\Omega |\nabla H|^2,$$

and hence we have from (6.20) that

$$\left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\partial D_0)} + \left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\partial D)} \le C(l(\gamma_0^\epsilon \cup E) + l(\gamma^\epsilon \cup E)) \int_\Omega |\nabla H|^2, \tag{6.21}$$

where $l(\gamma_0^\epsilon)$ and $l(\gamma^\epsilon)$ denote the length of γ_0^ϵ and γ^ϵ.

Let $\Omega_1 = \{x \in \Omega | \text{dist}(x, \partial\Omega) > \delta_0\}$. Since ∇H is harmonic, one can apply the same argument of the harmonic measure used in the proof of Proposition 3.4 and the standard interior estimate to see that there exist $C > 0$ and $0 < \alpha < 1$ such that for every $x \in \Omega_0$,

$$\begin{aligned}|\nabla H(x)|^2 &\le C \sup_{y \in B_{r_0/2}(0)} |\nabla H(y)|^{2\alpha} \sup_{y \in \Omega_1} |\nabla H(y)|^{2(1-\alpha)} \\ &\le C\left(\int_{D_0} |\nabla H(y)|^2\right)^\alpha \left(\int_\Omega |\nabla H(y)|^2\right)^{1-\alpha}.\end{aligned}$$

By integrating the left-hand side of the above inequality over Ω_0, we obtain

$$C\frac{\left(\int_{\Omega_0} |\nabla H(y)|^2\right)^{1/\alpha}}{\left(\int_\Omega |\nabla H(y)|^2\right)^{1/\alpha - 1}} \le \int_{D_0} |\nabla H(y)|^2 \le C\left\|\frac{\partial H}{\partial \nu}\right\|^2_{L^2(\partial D_0)}.$$

It then follows from (6.21) that

$$l(\gamma_0^\epsilon \cup E) + l(\gamma^\epsilon \cup E) \ge CA, \qquad A := \left(\frac{\int_{\Omega_0} |\nabla H(y)|^2}{\int_\Omega |\nabla H(y)|^2}\right)^{1/\alpha}. \tag{6.22}$$

By the definition of the ϵ-perturbation of a disk, there is a positive constant C so that

$$\begin{aligned}&l(\{x \in \partial B_r(a) : x \in B_{r_0+\epsilon}(0) \setminus B_{r_0-\epsilon}(0)\}) \\ &\qquad + l(\{x \in \partial B_{r_0}(0) : x \in B_{r+\epsilon}(a) \setminus B_{r-\epsilon}(a)\}) \ge CA.\end{aligned}$$

From the elementary geometry of two circles, there is a positive constant C depending only on r, r_0, A so that

$$|B_{r_0}(0) \setminus B_r(a)| + |B_r(a) \setminus B_{r_0}(0)| \leq C\epsilon.$$

It follows from the definition of the ϵ-perturbation of a disk that

$$|D_0 \setminus \overline{D}| + |D \setminus \overline{D_0}| \leq C\epsilon.$$

Now let us consider the remaining two cases:

(1) $\Omega \setminus \overline{D_0 \cup D}$ is connected and $\partial(\overline{D_0 \cup D}) \cap \partial(D_0 \cap D) = \emptyset$;

(2) $\Omega \setminus \overline{D_0 \cup D}$ is not connected.

From elementary geometry, it is not difficult to see that the above two cases occur when

$$|a| = r_0 + r + O(\epsilon). \tag{6.23}$$

Indeed, in case (1) there is a point $P \in \partial D \cap \partial D_0$ so that the outer normal vector of ∂D at P is in the opposite direction of the normal vector of ∂D_0 at P. Then (6.23) follows from the property of two circles and the ϵ-perturbation of a disk. In case (2), there is a point $Q = (Q_1, Q_2)$ which lies in a bounded component of $\mathbb{R}^2 \setminus \overline{D \cup D_0}$. Assume $Q_2 \geq 0$. The upper half vertical line $L := \{(x_1, x_2) : x_1 = Q_1 \text{ and } x_2 > Q_2\}$ starting from Q cuts or intersects one of the domains D and D_0. Assume that the vertical line L cuts D. It follows from Rolle's theorem that there is a point $P \in \partial D$ with $Q_2 \leq P_2$ such that the normal vector $\nu(P)$ at ∂D is parallel to the x_1-axis. Then $0 \leq Q_2 < P_2 = O(\epsilon)$, because ∂D is an ϵ-perturbation of a disk. This gives (6.23) because $|Q| > r_0 - \epsilon$ and $|Q - a| > r - \epsilon$.

Let $I_0 := \partial D_0 \cap B_{r+\epsilon}(a)$ and $I := \partial D \cap B_{r_0+\epsilon}(0)$. Then it is easy to see that

$$l(I_0 \cup I) \leq C\sqrt{\epsilon}. \tag{6.24}$$

As in (6.8), we apply the Rellich identity with vector field x and we obtain

$$\int_{\partial(D_0 \setminus \overline{D})} \langle x, \nu \rangle \left| \frac{\partial U^\#}{\partial \nu} \right|^2 = \int_{\partial(D_0 \setminus \overline{D})} \langle x, \nu \rangle \left| \frac{\partial U}{\partial T} \right|^2 - 2 \int_{\partial(D_0 \setminus \overline{D})} \langle x, T \rangle \frac{\partial U}{\partial T} \frac{\partial U^\#}{\partial \nu}. \tag{6.25}$$

(Note that $\mathcal{R} = 0$ with the vector field x.) Similarly, we obtain

(6.26)

$$\int_{\partial(D \setminus \overline{D_0})} \langle x - a, \nu \rangle \left| \frac{\partial U^\#}{\partial \nu} \right|^2 = \int_{\partial(D \setminus \overline{D_0})} \langle x - a, \nu \rangle \left| \frac{\partial U}{\partial T} \right|^2 - 2 \int_{\partial(D \setminus \overline{D_0})} \langle x - a, T \rangle \frac{\partial U}{\partial T} \frac{\partial U^\#}{\partial \nu}.$$

Since $\frac{\partial U}{\partial T} = 0$ on $(\partial D \cup \partial D_0) \setminus (I \cup I_0)$, it follows from (6.25) and (6.26) that

$$\int_{(\partial D \cup \partial D_0) \setminus (I_0 \cup I)} \left| \frac{\partial U^\#}{\partial \nu} \right|^2 \leq C \int_{I_0 \cup I} \left| \nabla U^\# \right|^2.$$

From Lemma 6.4 and the above estimate, we obtain as before that

$$\|\varphi_0\|_{L^2(\partial D_0)} + \|\varphi\|_{L^2(\partial D)} \leq C \left(\|\varphi_0\|_{L^2(I_0)} + \|\varphi\|_{L^2(I)} \right),$$

which is similar to (6.19). We repeat the argument as before and obtain $l(I \cup I_0) > C$ as in (6.22). This is not possible for small ϵ because of (6.24). This completes the proof. □

Acknowledgments. This work was completed before professor E. Fabes passed away. We will remember him as a man of humanism and as a great teacher in our life. We would like to thank the referee for several helpful comments.

REFERENCES

[AIP] G. Alessandrini, V. Isakov, and J. Powell, *Local uniqueness in the inverse problem with one measurement*, Trans. Amer. Math. Soc., 347 (1995), pp. 3031–3041.

[BF] H. Bellout and A. Friedman, *Identification problem in potential theory*, Arch. Rational Mech. Anal., 101 (1988), pp. 143–160.

[BFI] H. Bellout, A. Friedman, and V. Isakov, *Inverse problem in potential theory*, Trans. Amer. Math. Soc., 332 (1992), pp. 271–296.

[BFS] B. Barcelo, E. Fabes, and J.K. Seo, *The inverse conductivity problem with one measurement: Uniqueness for convex polyhedra*, Proc. Amer. Math. Soc., 122 (1994), pp. 183–189.

[EFV] 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.

[F] G.B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.

[FI] A. Friedman and V. Isakov, *On the uniqueness in the inverse conductivity problem with one measurement*, Indiana Univ. Math. J., 38 (1989), pp. 553–580.

[FJR] E.B. Fabes, M. Jodeit, and N.M. Riviére, *Potential techniques for boundary value problems on C^1 domains*, Acta Math., 141 (1978), pp. 165–186.

[GT] D. Gilbarg and N.S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, New York, 1983.

[IP] V. Isakov and J. Powell, *On the inverse conductivity problem with one measurement*, Inverse Problems, 6 (1990), pp. 311–318.

[KS1] H. Kang and J.K. Seo, *Layer potential technique for the inverse conductivity problem*, Inverse Problems, 12 (1996), pp. 267–278.

[KS2] H. Kang and J.K. Seo, *Inverse conductivity problem with one measurement: Uniqueness of balls in $\mathbb{R}^3$*, SIAM J. Appl. Math., to appear.

[KSS1] H. Kang, J.K. Seo, and D. Sheen, *The inverse conductivity problem with one measurement: Stability and estimation of size*, SIAM J. Math. Anal., 28 (1997), pp. 1389–1405.

[KSS2] H. Kang, J.K. Seo, and D. Sheen, *Numerical identification of discontinuous conductivity coefficients*, Inverse Problems, 13 (1997), pp. 113–123.

[S] J.K. Seo, *A uniqueness result on inverse conductivity problem with two measurements*, J. Fourier Anal. Appl., 2 (1996), pp. 227–235.

[V] G.C. Verchota, *Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 721–746

LOWER BOUNDS FOR GENERALIZED GINZBURG–LANDAU FUNCTIONALS*

ROBERT L. JERRARD†

Abstract. We study properties of Ginzburg–Landau functionals $I_U^\epsilon(\cdot)$, defined for functions $u \in W^{1,n}(U;\mathcal{R}^n)$, where $U \subset \mathcal{R}^n$. In particular, we establish lower bounds relating the energy $I_U^\epsilon(u)$ to the Brouwer degree of u, and we prove under additional hypotheses that the energy concentrates on a small number of small sets. As a consequence we deduce some compactness theorems. Such estimates are useful in studying Ginzburg–Landau-type PDEs associated with the functional I_U^ϵ.

Key words. Ginzburg–Landau functional, lower bounds, energy concentration, compactness

AMS subject classifications. 35J50, 35Q80

PII. S0036141097300581

1. Introduction.

We study Ginzburg–Landau functionals of the form

$$I^\epsilon(u) := \int_U \mathcal{E}^\epsilon[u]dx,$$

where U is a bounded open subset of $\mathcal{R}^n$, $u \in W^{1,n}(U;\mathcal{R}^n)$, and

$$\mathcal{E}^\epsilon[u] := \frac{1}{n}|\nabla u|^n + \frac{1}{4\epsilon^2}(1-|u|^2)^2.$$

Our results apply also to the functional with magnetic potential,

$$I^\epsilon_{\rm mag}(u,A) := \int_U \mathcal{E}^\epsilon_{\rm mag}[u,A]dx,$$

where U is a bounded open subset of $\mathcal{R}^2$ and

$$\mathcal{E}^\epsilon_{\rm mag}[u,A] := \frac{1}{2}|\nabla_A u|^2 + \frac{1}{2}|\nabla \times A|^2 + \frac{1}{4\epsilon^2}(1-|u|^2)^2.$$

The notation is explained below.

When there is no possibility of confusion, we will typically write $\mathcal{E}^\epsilon$ and $\mathcal{E}^\epsilon_{\rm mag}$, suppressing the dependence on u and A.

It is known that if the trace of u on ∂U is a fixed function of nonzero Brouwer degree, then $I^\epsilon(u) > C\ln(\frac{1}{\epsilon})$, for ϵ small, and similarly for $I^\epsilon_{\rm mag}(u,A)$. We give a new proof of lower bounds of this form, and we show that under additional hypotheses the unbounded part of the energy is concentrated on a small number of small sets. As a consequence we establish new results on the weak compactness of functions with nonzero degree.

In the magnetic case, we further show that if the energy of a pair (u,A) is within $O(1)$ of the appropriate lower bound, then the term $\int |\nabla \times A|^2$ is bounded independent of ϵ.

*Received by the editors June 30, 1997; accepted for publication (in revised form) June 4, 1998; published electronically April 20, 1999. The research of this author was partially supported by the Army Research Office and the National Science Foundation through the Center for Nonlinear Analysis and NSF grant DMS 96-00080.

http://www.siam.org/journals/sima/30-4/30058.html

†Department of Mathematics, University of Illinois, 1409 W. Green St., Urbana, IL 61801 (rjerrard@math.uiuc.edu).

The first results of this character were proven by Bethuel, Brezis, and Hélein [2], using lower bounds established by Brezis, Merle, and Rivière [4]. Other important contributions include papers of Struwe [11], Bethuel and Rivière [3], Han and Li [6], and Hong [7]. Most of these results deal only with sequences of energy-minimizers, and provide detailed information about their asymptotic behavior. A paper of Lin [8] establishes compactness results under weaker assumptions which are naturally satisfied by solutions of parabolic equations.

Our techniques are substantially new, and for many purposes they are more precise and flexible than earlier arguments. In particular, they require no a priori control on the modulus of continuity of a function u. We thus do not need to use any regularity theory—we instead rely entirely on elementary arguments.

In particular, our results show that the compactness of energy-minimizing sequences does not in any way depend on the fact that they solve some PDE; instead, it is merely an outcome of the real variable structure of the Ginzburg–Landau functional.

After completing this paper, we have become aware of several related works, including a recent paper of Almeida and Bethuel [1] and preprints of Lin [9] and Sandier [10]. These papers deal mainly with the case $n = 2$; as far as we know, our compactness results are the only ones valid in higher dimensions. The main focus of [1] is on topological methods, but their work can be used to deduce the compactness of nonminimizing, though appropriately bounded, sequences. Lin [9] and Sandier [10] both establish results similar to ours. Sandier uses estimates on annuli, rather similar to ours, together with an elegant application of the coarea formula. Lin's arguments are different; he uses a regularization to reduce the general case (with no control on the modulus of continuity) to a situation he studied earlier in [8]. Lin further applies this result to the study of asymptotic vortex dynamics in solutions of a Ginzburg–Landau wave equation.

In general, results of the sort we establish are very useful in studying vortex dynamics for evolutionary equations of Ginzburg–Landau-type, especially in situations where there is only weak control over the modulus of continuity.

Ginzburg–Landau-type functionals in n-dimensions. Suppose that ∂U is smooth, and let $g : \partial U \to S^{n-1}$ be smooth, with

$$|\deg(g; \partial U)| = d.$$

We are interested in the case $d \neq 0$, and given this, we assume for convenience that $\deg(g; \partial U) > 0$.

Let

$$W^{1,n}_g(U; \mathcal{R}^n) := \{u \in W^{1,n}(U; \mathcal{R}^n) \,|\, u = g \text{ on } \partial U\}. \tag{1.1}$$

We define a constant

$$\kappa_n = \frac{1}{n}(n-1)^{n/2}|\partial B_1|. \tag{1.2}$$

The following result is proven in Han and Li [6] and Hong [7], who obtain it by analyzing energy-minimizing sequences. These authors obtain also much more detailed information about the limiting behavior of minimizers.

THEOREM 1.1. *If $u \in W^{1,n}_g(U; \mathcal{R}^n)$, then*

$$\int_U \mathcal{E}^\epsilon dx \geq d\kappa_n \ln\left(\frac{1}{\epsilon}\right) - C. \tag{1.3}$$

The constant here depends on the domain U and the data g.

We will show that a matching upper bound implies energy concentration. In fact, the following result states that energy concentrates around at most d points.

THEOREM 1.2. *Suppose that $u \in W_g^{1,n}(U;\mathcal{R}^n)$. There exists a number $\sigma_0 > 0$ (depending only on the domain U) such that, for any $\sigma \in (0,\sigma_0]$, if*

$$\int_U \mathcal{E}^\epsilon dx \le d\kappa_n \ln\left(\frac{1}{\epsilon}\right) + C \tag{1.4}$$

and ϵ is sufficiently small, i.e., $\epsilon \le \epsilon_0(C,\sigma)$, then there are points $\{x_1,\dots,x_m\}$ and positive integers $\{d_1,\dots,d_m\}$ such that

$$\sum_i d_i = d,$$

and

$$\int_{\cup_i B_{d_i\sigma}(x_i)\cap U} \mathcal{E}^\epsilon dx \ge d\kappa_n \ln\left(\frac{\sigma}{\epsilon}\right) - C.$$

Note that Theorem 1.2 implies Theorem 1.1. Indeed, if $u \in W_g^{1,n}(U;\mathcal{R}^n)$ satisfies (1.4), then (1.3) follows by Theorem 1.2, and if (1.4) does not hold, then (1.3) follows trivially.

A compactness result follows as an easy consequence of Theorem 1.2

THEOREM 1.3. *Suppose u^ϵ is a collection of functions in $W_g^{1,n}(U;\mathcal{R}^n)$ and that u^ϵ satisfies*

$$\int_U \mathcal{E}^\epsilon dx \le d\kappa_n \ln\left(\frac{1}{\epsilon}\right) + C$$

for every $\epsilon > 0$. Then there exist points $x_1,\dots,x_m \in \bar{U}$, with $m \le d$, a subsequence $\epsilon_k \to 0$, and a function $u \in W^{1,n}_{loc}(U \setminus \{x_1,\dots,x_m\}; S^{n-1})$ such that

$$u^{\epsilon_k} \rightharpoonup u \qquad \text{weakly in } W^{1,n}_{loc}(U \setminus \{x_1,\dots,x_m\};\mathcal{R}^n).$$

Also, there are integers $d_i > 0$ for $i = 1,\dots,m$ such that $\sum d_i = d$ and

$$\mu^\epsilon := |\ln\epsilon|^{-1}\mathcal{E}^\epsilon\, dx \to \kappa_n \sum_{i=1}^m d_i \delta_{x_i}$$

weakly as measures.

Remarks. 1. It is straightforward to construct examples of functions satisfying the hypotheses of Theorem 1.3, so the theorem is not vacuous. Such constructions are standard and may be found in [2] and [7], for example. In particular, Theorem 1.3 implies the compactness of a sequence of energy-minimizers.

2. In Lin [9] it is shown that $m = d$ and that all points x_i lie in the interior of U. These results are almost certainly true in higher dimensions as well, but we do not prove them here.

3. Although we do not assume continuity in the statements of the above theorems, in many of our proofs we work with continuous functions. The general results as stated here will follow from regularization arguments.

The gauge-invariant Ginzburg–Landau functional. Consider the gauge-invariant functional

$$I^{\epsilon}_{\mathrm{mag}}(u;A) := \int_U \frac{1}{2}|\nabla_A u|^2 + \frac{1}{2}|\nabla \times A|^2 + \frac{1}{4\epsilon^2}(1-|u|^2)^2 dx.$$

Here $U \subset \mathcal{R}^2$ is assumed open and bounded, with smooth boundary. We now think of u as taking values in the complex plane $\mathcal{C}$, and $A = A_1 dx_1 + A_2 dx_2$ is a 1-form with coefficients $A_i \in H^1(U)$. We will identify A with the function $(A_1, A_2) \in H^1(U;\mathcal{R}^2)$. We define $\nabla \times A := A_{2,x_1} - A_{1,x_2}$ and $\nabla_A u := (\nabla - iA)u$, where $i = \sqrt{-1}$.

A thorough description of asymptotic behavior of minimizers, subject to gauge-invariant Dirichlet conditions, has been carried out by Bethuel and Rivière [3], who in particular determine a renormalized energy which governs the location of limiting singular points.

We will prove the following.

THEOREM 1.4. *Suppose that $u \in H^1_g(U;\mathcal{C})$ and $A \in H^1(U;\mathcal{R}^2)$. There exists a number $\sigma_0 > 0$ (depending only on the domain U) such that, for any $\sigma \in (0,\sigma_0]$, if*

$$\int_U \mathcal{E}^{\epsilon}_{mag} dx \le d\pi \ln\left(\frac{1}{\epsilon}\right) + C$$

and $\epsilon \le \epsilon_0(C,\sigma)$, then

$$\int_U |\nabla \times A|^2 dx \le C, \tag{1.5}$$

and there are points $\{x_1,\dots,x_m\}$ and positive integers $\{d_1,\dots,d_m\}$ such that

$$\sum_i d_i = d,$$

and

$$\int_{\cup_i B_{d_i\sigma}(x_i)\cap U} \mathcal{E}^{\epsilon}_{mag} dx \ge d\kappa_n \ln\left(\frac{\sigma}{\epsilon}\right) - C.$$

Note in particular the *upper* bound (1.5).

This immediately yields a lower bound, analogous to Theorem 1.1. We obtain also a compactness result along the lines of Theorem 1.3.

Notation, and preliminary remarks about degree. We let $B_r(x)$ denote the *closed* ball $\{y \in \mathcal{R}^n \mid |x-y| \le r\}$.

Let ω_n be the volume of the unit ball in $\mathcal{R}^n$.

We recall some facts about degree. A good general reference for this and related material is Brezis and Nirenberg [5].

When we refer to degree $\deg(u;\partial V)$, we always mean the degree of u around the origin. Informally, this counts the number of points in $u^{-1}(0)\cap V$, with multiplicity. In two dimensions, $\deg(u;\partial V)$ is just the winding number of ∂V around the origin.

More precisely, let $u \in W^{1,n}(U;\mathcal{R}^n)$, and suppose that $V \subset U$ and that V is bounded, with smooth boundary. If ess $\inf_{\partial V}|u| > 0$, then the Brouwer degree of u is defined by

$$\deg(u;\partial V) = \int_V \eta(u) \det Du \, dx, \tag{1.6}$$

where $\eta \in C^\infty(\mathcal{R}^n)$ satisfies

$$\int \eta = 1, \qquad \eta \geq 0, \qquad \text{spt } \eta \subset \{y \in \mathcal{R}^n : |y| < \text{ess inf}_{\partial V} |u|\}.$$

The degree is an integer, and it is independent of the specific choice of η and thus well defined. It can also be defined by the formula

$$\deg(u; \partial V) = \frac{1}{|\partial B_1|} \int_{\partial V} \det \nabla_\tau v \, dH^{n-1}. \tag{1.7}$$

Here $v := \frac{u}{|u|}|_{\partial V}$ is a map $\partial V \to \partial B_1$. The differential of v is a linear map $T_x \partial V \to T_{v(x)} \partial B_1$, and as such it can be expressed as an $(n-1) \times (n-1)$ matrix, in terms of orthonormal bases for $T_x \partial V$ and $T_{v(x)} \partial B_1$, which inherit natural orientations from the ambient spaces. This matrix is denoted $\nabla_\tau v$. The right-hand side of (1.7) is well defined, since $\det \nabla_\tau v$ is independent of the specific choice of bases. Also, it follows from the trace theorem that, under the stated assumptions, $v \in W^{1,n-1}(\partial V; \partial B_1)$, so the integration makes sense.

We will use the fact that

$$\det \nabla_\tau v \leq (n-1)^{-\frac{n-1}{2}} |\nabla_\tau v|^{n-1}, \tag{1.8}$$

which follows immediately from the inequality of arithmetic and geometric means.

It will be convenient for our purposes to use an approximation to the degree, which will enable us to ignore "inessential" components of the zero set of u. We will define this approximation only for $u \in C \cap W^{1,n}(U; \mathcal{R}^n)$. Before giving the definition, we introduce some notation.

We let S denote the set on which $|u|$ is small,

$$S := \{x \in U | \ |u(x)| \leq 1/2\}. \tag{1.9}$$

If we assume that u is continuous, then the connected components of S are closed, and each component S_i of S has a well-defined degree, given by the definition (1.6). This degree is an integer even when ∂S_i is not smooth, as may be seen by approximating S_i by smooth sets.

For $u \in C \cap W^{1,n}(U; \mathcal{R}^n)$ we may thus define the *essential* part of S,

$$S_E := \cup\{\text{components } S_i \text{ of } S | \deg(u; \partial S_i) \neq 0\} \tag{1.10}$$

and the *negligible* part of S,

$$\begin{aligned} S_N &:= \cup\{\text{components } S_i \text{ of } S | \deg(u; \partial S_i) = 0\} \\ &= S \setminus S_E. \end{aligned} \tag{1.11}$$

For any subset $V \subset U$ such that $\partial V \cap S_E \neq \emptyset$, we define the approximate degree

$$\text{dg}(u; \partial V) := \sum \{\deg(u; \partial S_i) \mid \text{components } S_i \text{ of } S_E \text{ such that } S_i \subset\subset V\}. \tag{1.12}$$

The advantage of using the approximate degree dg is that it allows us to conduct our analysis as if every component of S has nonzero degree, which is useful in one or two places. Note that dg agrees with the ordinary degree on sets for which both are defined. Indeed, for many purposes the distinction between the two can be ignored with very little loss of understanding.

2. An estimate on spheres. In this section we establish an estimate on spheres. A similar estimate is proven by Han and Li [6] and Hong [7], who establish a lower bound for $||Du||_{L^n}$ on an annulus, under the assumption that $\int(1-|u|^2)^2dx$ is small. Both of these proofs use a good deal of machinery from elliptic regularity theory.

We use the following.

Notation. Recall that we have defined the constant $\kappa_n = \frac{1}{n}(n-1)^{n/2}|\partial B_1|$. We also define

$$\lambda^\epsilon(r;d) = \min_{m\in[0,1]}\left[m^n\frac{\kappa_n|d|^{\frac{n}{n-1}}}{r} + \frac{1}{C\epsilon}(1-m)^N\right], \tag{2.1}$$

where $C, N > 0$ will be specified below and will depend only on the dimension n.

The main result of this section is the following theorem.

THEOREM 2.1. *λ^ϵ has the following properties: First,*

$$\lambda^\epsilon(r;d) \geq \kappa_n|d|^{\frac{n}{n-1}}\frac{1}{r}\left(1 - C(d)\frac{\epsilon^\alpha}{r^\alpha}\right) \tag{2.2}$$

for $\alpha = \frac{1}{N-1} > 0$.

Second, if $r > \epsilon$, $u \in W^{1,n}(\partial B_r; B_1)$, and $|\deg(u;\partial B_r)| = d > 0$, then

$$\int_{\partial B_r}\mathcal{E}^\epsilon\, dH^{n-1} \quad \geq \quad \lambda^\epsilon(r;d). \tag{2.3}$$

We assume several lemmas, and use them to prove the following theorem.

Proof. 1. We first prove (2.3). Given u as stated, we define

$$\rho := |u|, \qquad v := u/\rho, \qquad m := 1 \wedge \min_{x\in\partial B_r}\rho(x). \tag{2.4}$$

Since $n \geq 2$, we see that

$$\begin{aligned}|\nabla_\tau u|^n &\geq |\nabla_\tau\rho|^n + \rho^n|\nabla_\tau v|^n\\ &\geq |\nabla_\tau\rho|^n + m^n|\nabla_\tau v|^n.\end{aligned}$$

Thus

$$\begin{aligned}\int_{\partial B_r}\mathcal{E}^\epsilon &\geq \int_{\partial B_r}\frac{m^n}{n}|\nabla_\tau v|^n + \frac{1}{n}|\nabla_\tau\rho|^n + \frac{1}{\epsilon^n}(1-\rho^2)^2\\ &\geq m^n\frac{\kappa_n|d|^{\frac{n}{n-1}}}{r} + \frac{1}{C\epsilon}|1-m|^N\end{aligned}$$

by Lemmas 2.3 and 2.4 below. This last inequality and the definition of λ^ϵ directly imply (2.3).

2. Next, for n, N as above, $m \in [0,1]$, and any fixed constant $K > 0$, we estimate

$$\begin{aligned}1 - m^n &\leq n(1-m)\\ &= \left((KN)^{1/N}(1-m)\right)\frac{n}{(KN)^{1/N}}\\ &\leq K(1-m)^N + \frac{N-1}{N}\left(\frac{n}{(KN)^{1/N}}\right)^{N/(N-1)}\\ &= K(1-m)^N + CK^{-1/(N-1)}\end{aligned}$$

using Young's inequality. Thus

$$\max_{m\in[0,1]}\left[(1-m^n)-K(1-m)^N\right]\le CK^{-1/(N-1)} \tag{2.5}$$

for any $K>0$.

It follows that for any $A,B>0$

$$\begin{aligned}\min_{m\in[0,1]}\left[m^nA+(1-m)^NB\right] &= A-A\max_{m\in[0,1]}\left[(1-m^n)-\frac{B}{A}(1-m)^N\right]\\ &\ge A\left(1-C\left(\frac{A}{B}\right)^{1/(1-N)}\right).\end{aligned}$$

This immediately gives (2.2). □

We now fill in the proofs of the lemmas used above. We continue to assume the hypotheses of Theorem 2.1, so that, for example, $u\in W^{1,n}(\partial B_r;\partial B_1)$. We first recall the following estimate of Morrey.

LEMMA 2.2. *ρ is Hölder continuous on ∂B_r, and in fact*

$$|\rho(x)-\rho(y)|\le C||\nabla_\tau\rho||_{L^n}|x-y|^{1/n}$$

for some constant C independent of r.

Proof. The stated estimate is invariant under rescalings, so it suffices to prove it on the unit sphere. This, however, follows easily from the standard Morrey inequality on $\mathcal{R}^{n-1}$. □

LEMMA 2.3. *If $r\ge\epsilon$, then*

$$\int_{\partial B_r}\frac{1}{n}|\nabla_\tau\rho|^n+\frac{1}{\epsilon^n}(1-\rho^2)^2dH^{n-1}\ge\frac{1}{C\epsilon}|1-m|^N$$

for some $C,N>0$.

Remark. It is clear from the proof that this lemma does not depend on the exact form of $W(\rho):=(1-\rho^2)^2$.

Proof. Let

$$\gamma:=\int_{\partial B_r}\frac{1}{n}|\nabla_\tau\rho|^ndH^{n-1},$$

and let $x_{\min}\in\partial B_r$ be a point at which $\rho(x_{\min})=m$.

Lemma 2.2 implies that

$$\begin{aligned}\rho(x)&\le m+C\gamma^{1/n}|x_{\min}-x|^{1/n}\\ &\le\frac{1+m}{2}\qquad\text{whenever } |x-x_{\min}|\le\frac{|1-m|^n}{C\gamma}.\end{aligned} \tag{2.6}$$

Since $r\ge\epsilon$ and $x_{\min}\in\partial B_r$,

$$H^{n-1}\left(\partial B_r\cap B_\sigma(x_{\min})\right)\ge C^{-1}(\sigma^{n-1}\wedge\epsilon^{n-1}) \tag{2.7}$$

for any $\sigma>0$. Since $(1-\rho^2)^2\ge C^{-1}|1-m|^2$ whenever $\rho\le(1+m)/2$, we deduce from (2.6) and (2.7) that

$$\int_{\partial B_r}(1-\rho^2)^2dH^{n-1}\ \ge\ C^{-1}|1-m|^2\left(\epsilon^{n-1}\wedge\frac{|1-m|^{n(n-1)}}{\gamma^{n-1}}\right).$$

It follows that

$$\int_{\partial B_r} \frac{1}{n}|\nabla_\tau \rho|^n + \frac{1}{\epsilon^n}(1-\rho^2)^2 dH^{n-1} \;\geq\; \gamma + \frac{1}{C\epsilon^n}|1-m|^2\left(\epsilon^{n-1} \wedge \frac{|1-m|^{n(n-1)}}{\gamma^{n-1}}\right).$$

The conclusion of the lemma is obvious if $\epsilon^{n-1} \leq |1-m|^{n(n-1)}\gamma^{1-n}$. If the other inequality holds, the desired estimate follows by using calculus to minimize over $\gamma > 0$. □

Our next assertion seems to be quite well known; it appears without proof in Han and Li [6], who remark that it is elementary. Nevertheless, we provide a proof for the reader's convenience.

LEMMA 2.4. *Suppose that $v \in W^{1,n}(\partial B_r, \partial B_1)$ for some $r > 0$ and that*

$$\deg(v; \partial B_r) = d.$$

Then

$$\int_{\partial B_r} \frac{1}{n}|\nabla_\tau v|^n dH^{n-1} \geq \frac{\kappa_n}{r}|d|^{\frac{n}{n-1}}.$$

Proof. We compute, using (1.8) in the second line below,

$$\begin{aligned}
\deg(v; \partial B_r) &= \frac{1}{|\partial B_1|}\int_{\partial B_r} \det \nabla_\tau v \; dH^{n-1} \\
&\leq \frac{1}{|\partial B_1|}(n-1)^{-\frac{n-1}{2}} \int_{\partial B_r} |\nabla_\tau v|^{n-1} dH^{n-1} \\
&\leq \frac{(n-1)^{-\frac{n-1}{2}}}{|\partial B_1|}\left(\int_{|\partial B_r|} |\nabla_\tau v|^n dH^{n-1}\right)^{\frac{n-1}{n}} |\partial B_r|^{\frac{1}{n}}.
\end{aligned}$$

After some rearranging, this becomes the conclusion of the lemma. □

Finally, we record a technical refinement of Lemma 2.3 which will be needed later.

LEMMA 2.5. *Suppose that $U \subset \mathcal{R}^n$ is a bounded open subset with a smooth boundary. Then there exists $\hat{r} > 0$ such that for every $r \in [\epsilon, \hat{r})$ and every $y \in U$ we have*

$$\int_{\partial B_r(y)\cap U} |\nabla_\tau \rho|^n + \frac{1}{\epsilon^n}(1-\rho^2)^2 dH^{n-1} \geq \frac{1}{C\epsilon}|1-m|^N,$$

where C depends on the smoothness of the boundary and on the dimension n.

Proof. The proof follows that of Lemma 2.3, except that in place of (2.7) we substitute

$$H^{n-1}\left(\partial B_r(y) \cap U \cap B_\sigma(x_{\min})\right) \geq C^{-1}(\sigma^{n-1} \wedge r^{n-1}) \tag{2.8}$$

for $y \in U, x_{\min} \in \partial B_r(y) \cap U$. Since $r > \epsilon$ by hypothesis, this is enough to extend the earlier argument to the present case.

We deduce (2.8) as a consequence of the smoothness of ∂U. Indeed, (2.8) is clear if U is the half-space $\{x \in \mathcal{R}^n : x_n > 0\}$. The general case follows by flattening out the boundary. □

3. Properties of Λ^ϵ. In this section we define a function Λ^ϵ, which provides a convenient way of keeping track of lower bounds on balls, and we record several properties of Λ^ϵ.

We define

$$\Lambda^\epsilon(s) := \int_0^s \lambda^\epsilon(r;1) \wedge \frac{c_0}{\epsilon} dr, \tag{3.1}$$

where c_0 is a constant to be selected below, depending only on the dimension n.

We first note some elementary properties of Λ^ϵ.

PROPOSITION 3.1. *$\Lambda^\epsilon(\cdot)$ is increasing, and moreover,*

$$\Lambda^\epsilon(r+s) \le \Lambda^\epsilon(r) + \Lambda^\epsilon(s) \qquad \forall\, r, s \ge 0, \tag{3.2}$$

$$s \mapsto \frac{1}{s}\Lambda^\epsilon(s) \qquad \textit{is nonincreasing,} \qquad \textit{and} \tag{3.3}$$

$$\Lambda^\epsilon(r) \ge \kappa_n \ln\left(\frac{r}{\epsilon}\right) - C(n) \qquad \forall\, r \ge 0. \tag{3.4}$$

Proof. From the definition (2.1) of λ^ϵ, it is clear that $\lambda^\epsilon > 0$ and that $r \mapsto \lambda^\epsilon(r;1)$ is nonincreasing. The first of these facts implies that Λ^ϵ is increasing; from the two facts together it is easy to see that (3.2) holds.

Next, (3.3) is clear, since $\frac{1}{s}\Lambda^\epsilon(s)$ is just the average over the interval $[0,s]$ of the nonincreasing function $r \mapsto \lambda^\epsilon(r;1) \wedge \frac{c_0}{\epsilon}$.

To prove the lower bound (3.4), recall first from Theorem 2.1 that

$$\lambda^\epsilon(r,1) \ge \kappa_n \left(\frac{1}{r} - \frac{C\epsilon^\alpha}{r^{1+\alpha}}\right).$$

As a result, $\lambda^\epsilon \wedge \frac{c_0}{\epsilon} \ge \kappa_n(\frac{1}{r} - \frac{C\epsilon^\alpha}{r^{1+\alpha}})$ whenever $r \ge c_1\epsilon$ for some $c_1 > 0$. Thus

$$\begin{aligned}
\Lambda^\epsilon(r) &\ge \int_{c_1\epsilon}^r \kappa_n \left(\frac{1}{s} - \frac{C\epsilon^\alpha}{s^{1+\alpha}}\right) ds \\
&\ge \kappa_n \ln\left(\frac{r}{c_1\epsilon}\right) - C\epsilon^\alpha \int_{c_1\epsilon}^\infty s^{-1-\alpha}\, ds \\
&\ge \kappa_n \ln\left(\frac{r}{\epsilon}\right) - C. \qquad \square
\end{aligned}$$

We next restate the lower bounds of the previous sections in terms of Λ^ϵ.

PROPOSITION 3.2. *Let $u \in C \cap W^{1,n}(U;\mathcal{R}^n)$, $\epsilon \le r_0 \le r_1$, and suppose that $|\mathrm{dg}\,(u;\partial B_\rho)| = d > 0$ $\forall$ $\rho \in [r_0, r_1]$, then*

$$\int_{B_{r_1}\setminus B_{r_0}} \mathcal{E}^\epsilon dx \ge d\left[\Lambda^\epsilon\left(\frac{r_1}{d}\right) - \Lambda^\epsilon\left(\frac{r_0}{d}\right)\right]. \tag{3.5}$$

Remark. Recall that dg is defined in (1.12). Stating this estimate in terms of dg allows us to ignore the "negligible" set S_N in our later arguments.

Proof

1. We first claim that if $|\mathrm{dg}(u;\partial B_r)| = d > 0$, then

$$\int_{\partial B_r} \mathcal{E}^\epsilon \, dH^{n-1} \quad \geq \quad \lambda^\epsilon(r;d) \wedge \frac{c_0}{\epsilon}$$

for some $c_0(n)$.

To see this, observe that the definition of dg implies that either $\deg(u;\partial B_r) = d$, in which case the claim follows immediately from (2.3), or $\partial B_r \cap S_N \neq \emptyset$, which implies that $m \leq \frac{1}{2}$, in the notation (2.4). Lemma 2.3 then implies that

$$\int_{\partial B_r} \mathcal{E}^\epsilon dH^{n-1} \geq \frac{c_0}{\epsilon}$$

if c_0 is small enough.

2. From the definition (2.1) of λ^ϵ it is clear that $\lambda^\epsilon(r;d) \geq \lambda^\epsilon(\frac{r}{d};1)\ \forall\ r > 0, d \geq 1$. We use this fact and step 1 to compute

$$\begin{aligned}
\int_{B_{r_1}\setminus B_{r_0}} \mathcal{E}^\epsilon dx &= \int_{r_0}^{r_1} \int_{\partial B_r} \mathcal{E}^\epsilon dH^{n-1} dr \\
&\geq \int_{r_0}^{r_1} \lambda^\epsilon\left(\frac{r}{d};1\right) \wedge \frac{c_0}{\epsilon} dr \\
&= d\int_{\frac{r_0}{d}}^{\frac{r_1}{d}} \lambda^\epsilon(r;1) \wedge \frac{c_0}{\epsilon} dr \quad = \quad d\left[\Lambda^\epsilon\left(\frac{r_1}{d}\right) - \Lambda^\epsilon\left(\frac{r_0}{d}\right)\right]. \qquad \square
\end{aligned}$$

The final property of Λ that we will need is the following.

PROPOSITION 3.3. *Suppose that $u \in W^{1,n}(U;\mathcal{R}^n)$ and that u is continuous. If $S_E \subset\subset U$, then there is a collection of closed, pairwise disjoint balls $\{B_i\}_{i=1}^k$ with radii r_i such that*

$$S_E \subset \cup_{i=1}^k B_i, \tag{3.6}$$

$$r_i \geq \epsilon \qquad \forall\, i; \tag{3.7}$$

$$B_i \cap S_E \neq \emptyset \qquad \text{for each } i, \tag{3.8}$$

$$\int_{B_i \cap U} \mathcal{E}^\epsilon dx \ \geq\ \frac{c_0}{\epsilon} r_i \ \geq\ \Lambda^\epsilon(r_i). \tag{3.9}$$

Remarks. 1. It is not true in general that the entire zero set of u can be covered by balls satisfying a lower bound of the form (3.9). This is the reason we need to decompose the set S into S_E and S_N (see (1.10) and (1.11)) and to introduce the approximate degree dg.

2. The technical assumption $S_E \subset\subset U$ is satisfied if $|u| > 1/2$ on ∂U.

We start by proving several lemmas. The first is a trivial observation which we will use repeatedly.

LEMMA 3.1. *Given any finite collection of closed balls in* $\mathbf{R}^k$, *say* $\{B_i\}_{i=1}^N$, *we can find a collection* $\{\tilde{B}_i\}_{i=1}^{\tilde{N}}$ *of pairwise disjoint balls such that*

$$\bigcup_{i=1}^{N} B_i \subset \bigcup_{i=1}^{\tilde{N}} \tilde{B}_i,$$

$$\sum_{B_j \subset \tilde{B}_i} \operatorname{diam} B_j = \operatorname{diam} \tilde{B}_i, \tag{3.10}$$

$$\tilde{N} \leq N, \quad \textit{with strict inequality unless } \{B_i\}_{i=1}^N \textit{ is pairwise disjoint.}$$

Proof. Replace pairs of intersecting balls B_i, B_j by larger single balls $\tilde{B}$ such that $B_i \cup B_j \subset \tilde{B}$ and diam $\tilde{B}$ = diam B_i + diam B_j, continuing until a pairwise disjoint collection is reached. This collection has the stated properties. □

LEMMA 3.2. *Let* S_i *be a connected component of* S_E, *and assume that* $S_i \subset\subset U$. *Then*

$$\int_{S_i} |Du|^n dx \geq C^{-1} |\deg(u; \partial S_i)|. \tag{3.11}$$

Proof. Fix a nonnegative function $\eta \in C^\infty(\mathcal{R}^n)$ such that $\int \eta = 1$ and spt $\eta \subset B_{1/4}$. Then, by the definition (1.6) of degree,

$$\begin{aligned} |\deg(u; \partial S_i)| &= \left| \int_{S_i} \eta(u) \det Du \, dx \right| \\ &\leq C \int_{S_i} |\det Du| \, dx \\ &\leq C \int_{S_i} |Du|^n dx. \end{aligned}$$

In the last line we have used the inequality of arithmetic and geometric means. □

LEMMA 3.3. *Suppose that* $u \in W^{1,n}(U; \mathcal{R}^n)$ *and that* u *is continuous. If* $S_E \subset\subset U$, *then there is a collection of closed, pairwise disjoint sets* $\{C_i\}_{i=1}^k$ *such that*

$$S_E \subset \cup_{i=1}^k C_i, \tag{3.12}$$

$$\int_{C_i \cap U} \mathcal{E}^\epsilon dx \geq \frac{c_0}{\epsilon} \, (\operatorname{diam} C_i \vee \epsilon). \tag{3.13}$$

Proof. 1. First note that Lemma 3.2 implies that S_E has a finite number of components, and clearly each component has nonempty interior.

For each component $S_i \subset S_E$, select a point $x_i \in S_i$ and define

$$\rho_i := \sup\{r > 0 | \partial B_s(x_i) \cap S_E \neq \emptyset \quad \forall \, s \in (0, r)\}. \tag{3.14}$$

Clearly $S_i \subset B_{\rho_i}(x_i)$, so $S_E \subset \cup_i B_{\rho_i}(x_i)$. We assert also that if $c_1 = c_1(n, U)$ is sufficiently small, then

$$\int_{B_{\rho_i}(x_i) \cap U} \mathcal{E}^\epsilon dx \geq \frac{1}{\epsilon} c_1 \rho_i. \tag{3.15}$$

This follows from Lemma 3.2 if $\rho_i \le 2\epsilon$. If $\rho_i > 2\epsilon$, then Lemma 2.5 implies that

$$\begin{aligned}\int_{B_{\rho_i}(x_i)\cap U} \mathcal{E}^\epsilon dx &\ge \int_\epsilon^{\rho_i\wedge\hat r}\int_{\partial B_r\cap U} \mathcal{E}^\epsilon dH^{n-1}dr\\ &\ge \frac{1}{C\epsilon}((\rho_i\wedge\hat r)-\epsilon).\end{aligned}$$

Recall that $\hat r$ is some constant depending only on the geometry of U. If $2\epsilon \le \rho_i \le \hat r$, then $\rho_i - \epsilon \ge \frac{1}{2}\rho_i$ and (3.15) follows. If $2\epsilon \le \hat r < \rho_i$, then $((\rho_i\wedge\hat r)-\epsilon) \ge \frac{\hat r}{2}$. Moreover, $\rho_i \le \operatorname{diam} U \le C\hat r$, for some C depending on the geometry of U; thus (3.15) holds for appropriately small c_1.

2. Next, if $x_1 \in B_{\rho_j}(x_j)$ for some $j \ne 1$, the choice of ρ_j implies that $S_1 \subset B_{\rho_j}(x_j)$. If this holds, we may drop the ball $B_{\rho_1}(x_1)$ from the collection, and the remaining balls will still cover S_E. Proceeding in this fashion and relabelling as necessary, we obtain a collection $\{B_{\rho_i}(x_i)\}_{i=1}^k$ covering S_E, each satisfying (3.15) and such that

$$x_i \notin B_{\rho_j}(x_j) \qquad \text{whenever } i \ne j. \tag{3.16}$$

3. Let $C_i, i = 1,\dots,\tilde k$ be the connected components of $\cup_{i=1}^k B_{\rho_i}(x_i)$. We will complete the proof by showing that these sets satisfy (3.13).

Fix some i and define

$$\mathcal{B} := \{B_{\rho_j}(x_j) \mid B_{\rho_j}(x_j) \subset C_i\}.$$

According to the Besicovitch covering lemma, we may find subcollections $\mathcal{B}_1,\dots,\mathcal{B}_N$ such that the balls within each subcollection are pairwise disjoint, and the set of centers of balls in the original collection $\mathcal{B}$ is contained in the union of all the balls in all the subcollections. The latter fact, together with (3.16), implies that

$$\mathcal{B} = \cup_{l=1}^N \mathcal{B}_l.$$

Here N is an absolute constant depending only on the dimension n. So

$$\begin{aligned}N\int_{C_i\cap U}\mathcal{E}^e dx &\ge \sum_{l=1}^N \int_{(\cup_{\mathcal{B}_l} B_{\rho_j}(x_j))\cap U}\mathcal{E}^\epsilon dx\\ &= \sum_{\mathcal{B}}\int_{B_{\rho_j}(x_j)\cap U}\mathcal{E}^\epsilon dx\\ &\ge \sum_{\mathcal{B}}\frac{1}{\epsilon}c_1\rho_j\\ &\ge \frac{c_1}{\epsilon}\operatorname{diam} C_i.\end{aligned}$$

Also, each set C_i contains at least one component S_i of S_E by construction, so it is clear from Lemma 3.2 that

$$\int_{C_i\cap U}\mathcal{E}^e dx \ge C^{-1}$$

for each i. The proof of (3.13) is concluded by selecting c_0 sufficiently small. □

We can now easily complete the proof of Proposition 3.3.

Proof. 1. For each of the sets C_i found above, fix $x_i \in C_i$ and let

$$\rho_i := \sup\{r > 0 | \partial B_s(x_i) \cap C_i \neq \emptyset \quad \forall\, s \in (0,r)\} \;\leq\; \operatorname{diam} C_i;$$

$$\sigma_i := \epsilon \vee \rho_i \;\leq\; \epsilon \vee \operatorname{diam} C_i.$$

Let $\{B_i^{\text{new}}\}_{i=1}^k$ be the collection of pairwise disjoint balls formed by combining the balls $\{B_{\sigma_i}(x_i)\}$ following the algorithm of Lemma 3.1, and let r_i^{new} be the radius of B_i^{new}. Clearly this collection of balls satisfies (3.6), (3.7), and (3.8). Moreover, since the sets C_j from Lemma 3.3 are pairwise disjoint,

$$\begin{aligned}
\int_{B_i^{\text{new}} \cap U} \mathcal{E}^\epsilon dx &\geq \sum_{\{j : B_{\sigma_j}(x_j) \subset B_i^{\text{new}}\}} \int_{C_j \cap U} \mathcal{E}^\epsilon dx \\
&\geq \sum_{\{j : B_{\sigma_j}(x_j) \subset B_i^{\text{new}}\}} \frac{c_0}{\epsilon}(\epsilon \vee \operatorname{diam} C_j) \\
&\geq \frac{c_0}{\epsilon} \sum_{\{j : B_{\sigma_j}(x_j) \subset B_i^{\text{new}}\}} \sigma_j \\
&= \frac{c_0 r_i^{\text{new}}}{\epsilon}.
\end{aligned}$$

We have used (3.13) and (3.10) in the above estimates.

Finally, it is immediate from the definition (3.1) of Λ^ϵ that $\frac{c_0 r}{\epsilon} \geq \Lambda^\epsilon(r)$. ∎

4. Lower bounds and concentration. In this section we prove Theorem 1.2.

We assume that U is a bounded, open subset of $\mathcal{R}^n$. We will in fact establish that the conclusion of Theorem 1.2 is valid under somewhat more general boundary conditions than are stated in the introduction. Throughout this section we will assume that $u \in W^{1,n}(U)$ satisfies

$$|u(x)| \geq \frac{1}{2} \qquad \text{in } \{x \in U | \operatorname{dist}(x, \partial U) \leq r\}, \tag{4.1}$$

$$|\deg(u; \partial U)| = d > 0, \tag{4.2}$$

and

$$\int_U \mathcal{E}^\epsilon[u] dx \leq d\kappa_n \ln\left(\frac{1}{\epsilon}\right) + C \tag{4.3}$$

for some $\epsilon > 0$.

Our first lemma shows that (4.1)–(4.2) are in fact more general than the boundary conditions stated in the introduction. This construction is standard.

LEMMA 4.1. *Let $u \in W_g^{1,n}(U; \mathcal{R}^n)$, where g is a smooth function of nonzero degree and $W_g^{1,n}$ is as defined in* (1.1). *Then u can be extended to a function $\tilde{u}$, defined on a set $\tilde{U} \supset U$ and satisfying* (4.1)–(4.2) *above. Also, if u satisfies* (4.3), *then so does its extension $\tilde{u}$.*

Proof. To see this, let $\tilde{U} := \{x \in \mathcal{R}^n | \operatorname{dist}(x; U) \leq r\}$ for some $r > 0$. We assume that r is small enough that each point in $\tilde{U} \setminus U$ has a unique closest point in ∂U and that U and $\tilde{U}$ have the same topology. We define $\tilde{u}(x) = u(x)$ for $x \in U$, and

$$\tilde{u}(x) = g(y) \qquad \text{for the unique } y \in \partial U \text{ such that } \operatorname{dist}(x, \partial U) = |x - y|$$

if $x \in \tilde{U} \setminus U$. One then immediately sees that $\tilde{u}$ satsifies (4.1) and (4.2) on $\tilde{U}$. Also,

$$\int_{\tilde{U}\setminus U} \mathcal{E}^\epsilon(\tilde{u})\, dx \le C,$$

which yields the final assertion of the lemma. □

Theorem 1.2 will follow directly from the next proposition and the above lemma.

PROPOSITION 4.1. *Assume that u is continuous and satisfies* (4.1)–(4.3). *Then, given any $\sigma \in (0, \frac{r}{4d})$, there exists some $\epsilon_0(\sigma) > 0$ such that, for each $\epsilon \in (0, \epsilon_0)$, we can find a collection of closed balls $\{B_i\}_{i=1}^m$ of radius r_i and degree d_i such that*

(4.4) *the interiors of the balls are pairwise disjoint,*

$$\int_{B_i\cap U} \mathcal{E}^\epsilon[u]dx \ge \frac{r_i}{s}\Lambda^\epsilon(s), \qquad \text{for every } i, \text{ where } s := \min_j r_j/|d_j|, \tag{4.5}$$

$$S_E \subset\subset \bigcup_{i=1}^m B_i, \qquad \text{and } B_i \cap S_E \neq \emptyset \text{ for every } i. \tag{4.6}$$

Moreover,

$$s \in \left[\frac{\sigma}{2}, \sigma\right] \tag{4.7}$$

$$d_i \ge 0 \qquad \forall\, i. \tag{4.8}$$

Proof. 1. We first claim that if ϵ is sufficiently small and $\{B_i\}$ is a collection of balls satisfying (4.4)–(4.7), then (4.8) holds.

Noting that

$$\frac{r_i}{s} \ge |d_i| \qquad \forall\, i, \tag{4.9}$$

we use (4.4), (4.5), and (3.4) to compute

$$\begin{aligned}
\int_U \mathcal{E}^\epsilon dx &\ge \sum_i \int_{B_i\cap U} \mathcal{E}^\epsilon dx \\
&\ge \sum_i \frac{r_i}{s}\Lambda^\epsilon(s) \\
&\ge \sum_i |d_i|\Lambda^\epsilon(s) \\
&\ge \left[\kappa_n \ln\left(\frac{s}{\epsilon}\right) - C\right] \sum_i |d_i|. \qquad (4.10)
\end{aligned}$$

If $d_i < 0$ for some i, then $\sum |d_i| \ge d + 2$, in which case (4.3) and (4.7) yield

$$\kappa_n d \ln\left(\frac{1}{\epsilon}\right) + C \;\ge\; (d+2)\left[\kappa_n \ln\left(\frac{\sigma}{2\epsilon}\right) - C\right].$$

This is impossible for ϵ sufficiently small.

It therefore suffices to find a collection satisfying (4.4)–(4.7).

2. Next, if $\{B_i\}$ is a collection of balls satisfying (4.4), (4.5), (4.6), and if

$$s := \min_i \frac{r_i}{|d_i|} \leq \sigma,$$

then $B_i \cap \partial U = \emptyset$ for each i.

Suppose, toward a contradiction, that $B_i \cap \partial U \neq \emptyset$ for some i. From (4.6) we deduce that B_i contains a zero of u. With (4.1) this implies that B_i has radius $r_i \geq r/2 \geq 2d\sigma$. Then

$$\begin{aligned} \int_{B_i \cap U} \mathcal{E}^\epsilon dx &\geq \frac{r_i}{s}\Lambda^\epsilon(s) \\ &\geq \frac{r_i}{\sigma}\Lambda^\epsilon(\sigma) \\ &\geq 2d\left[\kappa_n \ln\left(\frac{\sigma}{\epsilon}\right) - C\right] \end{aligned} \tag{4.11}$$

by (4.5), (3.3), and (3.4). If ϵ is sufficiently small, this contradicts (4.3).

3. Let $\{B_i\}$ be the collection of balls found in Proposition 3.3. This collection satisfies (4.4), (4.5), and (4.6) by construction.

We now claim that, for this particular collection of balls, $s \leq \sigma/2$ if ϵ is sufficiently small. To see this, use (4.3), (3.9), and (4.9) to estimate

$$\begin{aligned} d\kappa_n \ln\left(\frac{1}{\epsilon}\right) + C &\geq \sum_i \int_{B_i \cap U} \mathcal{E}^\epsilon dx \\ &\geq \frac{1}{\epsilon} c_0 \sum_i r_i \\ &\geq \frac{1}{\epsilon} c_0 s \sum_i |d_i|. \end{aligned}$$

Thus in fact $s = O(\epsilon |ln\epsilon|)$, which certainly implies our claim.

We now successively modify these balls in such a way that (4.4), (4.5), and (4.6) remain true, and eventually (4.7) is satisfied.

At each step some balls are changed. We use the notation B_i^{old} and B_i^{new} to distinguish the ith ball, before and after it is modified, and similarly $r_i^{\text{old}}, r_i^{\text{new}}$, etc. When there is no possibility of confusion, we omit the superscripts.

4. *Expansion:* Note that the collection of balls obtained above is disjoint by construction. We modify these balls as follows.

(i) Identify the subcollection of balls B_i satisfying

$$\frac{r_i}{|d_i|} = \min_j \frac{r_j}{|d_j|} = s.$$

We refer to these as minimizing balls. The other balls are called nonminimizing balls.

(ii) Expand the minimizing balls uniformly by leaving the centers fixed and letting the radii grow. This is done in such a way that the ratio $s = r_i/|d_i|$ is always uniform for the minimizing balls. Thus one may think of the expansion as simply increasing the parameter s. Leave the nonminimizing balls unchanged.

More precisely, for $s \geq s^{\text{old}}$ define

$$B_i^s := \begin{cases} B_i^{\text{old}} & \text{if } B_i^{\text{old}} \text{ is nonminimizing,} \\ B_{s|d_i^{\text{old}}|}(x_i^{\text{old}}) & \text{if } B_i^{\text{old}} \text{ is minimizing.} \end{cases}$$

(iii) Define

$$s^{\text{new}} := \sup\{s \geq s^{\text{old}} | \text{ conditions (4.12)–(4.14) below are all satisfied.}\}$$

The conditions alluded to are

$$s < \frac{\sigma}{2}, \tag{4.12}$$

$$s < \frac{r_j^{\text{old}}}{|d_j^{\text{old}}|} \qquad \text{for all nonminimizing balls } B_j^{\text{old}}, \tag{4.13}$$

$$B_i^s \cap B_j^s = \emptyset \qquad \text{whenever } i \neq j, \tag{4.14}$$

where B_i^s, B_j^s are the expanded balls.

(iv) Let $B_i^{\text{new}} := B_i^{s^{\text{new}}}$.

(v) Note that, by step 2, $B_i^{\text{new}} \cap \partial U = \emptyset \ \forall\ i$, because $s^{\text{new}} \leq \sigma$, by (4.12).

(vi) We will verify in step 5 below that (4.4)–(4.6) hold for this new collection.

(vii) It must be the case that s^{new} violates one of (4.12), (4.13), or (4.14).

(viii) If (4.12) does not hold, then $s^{\text{new}} = \sigma/2$ and we are finished.

(ix) If (4.13) does not hold, then $s^{\text{new}} = r_j^{\text{old}}/|d_j^{\text{old}}|$ for some j such that B_j^{old} was nonminimizing. In this case we add all such balls B_j to the collection of minimizing balls and then start another expansion step.

(x) If (4.14) does not hold, we proceed with an "amalgamation step," as described in step 6 below.

Note that each time an expansion step terminates with (4.13) being satisfied, the number of nonminimizing balls decreases. As the total number of balls is finite, this can happen only finitely many times in succession. Thus a series of expansion steps must lead always to either (4.12) or (4.14) being satisfied.

5. It is clear that the collection of balls produced by the above expansion algorithm satisfies (4.6). Moreover, (4.4) holds as a consequence of the stopping criterion (4.14). So we need only to verify (4.5).

First, consider some i such that B_i^{old} was nonminimizing, which implies that $B_i^{\text{new}} = B_i^{\text{old}}$. Then

$$\int_{B_i^{\text{new}}} \mathcal{E}^\epsilon dx = \int_{B_i^{\text{old}}} \mathcal{E}^\epsilon dx \geq \frac{r_i}{s^{\text{old}}} \Lambda^\epsilon(s^{\text{old}}) \geq \frac{r_i}{s^{\text{new}}} \Lambda^\epsilon(s^{\text{new}}).$$

We have used (3.3) and the fact that $s^{\text{new}} \geq s^{\text{old}}$, which is clear from the expansion procedure.

Next, suppose that B_i^{old} was minimizing. We temporarily drop the subscripts i. The expansion procedure is carried out in such a way that $B^{\text{new}} \setminus B^{\text{old}}$ is an annulus with center x, inner radius r^{old}, and outer radius r^{new}; moreover

$$[B_{r^{\text{new}}}(x) \setminus B_{r^{\text{old}}}(x)] \cap S_E = \emptyset.$$

This last fact follows from (4.6) and the stopping criterion (4.14). This implies that $d^{\text{old}} = d^{\text{new}} = \deg(u; \partial B_r)\ \forall\ r \in [r^{\text{old}}, r^{\text{new}}]$. Proposition 3.2 thus allows us to estimate

$$\begin{aligned}\int_{B^{\text{new}}} \mathcal{E}^\epsilon dx &= \int_{B^{\text{new}} \setminus B^{\text{old}}} \mathcal{E}^\epsilon dx + \int_{B^{\text{old}}} \mathcal{E}^\epsilon dx \\ &\geq |d^{\text{new}}| \left[\Lambda^\epsilon\left(\frac{r^{\text{new}}}{|d^{\text{new}}|}\right) - \Lambda^\epsilon\left(\frac{r^{\text{old}}}{|d^{\text{old}}|}\right)\right] + \frac{r^{\text{old}}}{s^{\text{old}}}\Lambda^\epsilon(s^{\text{old}}) \\ &= |d^{\text{new}}|\Lambda^\epsilon\left(\frac{r^{\text{new}}}{|d^{\text{new}}|}\right) \\ &= \frac{r^{\text{new}}}{s^{\text{new}}}\Lambda^\epsilon(s^{\text{new}}).\end{aligned}$$

In the above computation, we have repeatedly used the fact that $s = r/|d|$ for minimizing balls.

6. *Amalgamation:* Suppose now that we have a collection of balls $\{B_i\}_i$ satisfying (4.4)–(4.6), such that (4.7) does not hold and that $B_i \cap B_j \neq \emptyset$ for some $i \neq j$. Combine the balls following the procedure of Lemma 3.1 to obtain a new collection of pairwise disjoint balls $\{B_j^{\text{new}}\}_j$. It is clear that this collection satisfies (4.4) and (4.6). We now verify that (4.5) also holds.

First, for any B_j^{new}, we have

$$r_j^{\text{new}} = \sum_{B_i^{\text{old}} \subset B_j^{\text{new}}} r_i^{\text{old}}, \qquad d_j^{\text{new}} = \sum_{B_i^{\text{old}} \subset B_j^{\text{new}}} d_i^{\text{old}}.$$

It is thus clear that

$$s^{\text{new}} = \min_j \frac{r_j^{\text{new}}}{|d_j^{\text{new}}|} \geq s^{\text{old}}. \tag{4.15}$$

Also, for each j,

$$\begin{aligned}\int_{B_j^{\text{new}}} \mathcal{E}^\epsilon dx &\geq \sum_{B_i^{\text{old}} \subset B_j^{\text{new}}} \int_{B_i^{\text{old}}} \mathcal{E}^\epsilon dx \\ &\geq \frac{1}{s^{\text{old}}}\Lambda^\epsilon(s^{\text{old}}) \sum_{B_i^{\text{old}} \subset B_j^{\text{new}}} r_i^{\text{old}} \\ &\geq \frac{r_j^{\text{new}}}{s^{\text{new}}}\Lambda^\epsilon(s^{\text{new}})\end{aligned}$$

by (3.10), (3.3), and (4.15).

7. We next show that $s^{\text{new}} \leq \sigma$. Recall that we have assumed that $s^{\text{old}} \leq \sigma/2$. This implies that

$$\begin{aligned}\sum_i \int_{B_i^{\text{old}}} \mathcal{E}^\epsilon dx &\geq \sum_i \frac{r_i^{\text{old}}}{s^{\text{old}}}\Lambda^\epsilon(s^{\text{old}}) \\ &\geq \sum_i \frac{r_i^{\text{old}}}{\sigma/2}\Lambda^\epsilon\left(\frac{\sigma}{2}\right) \\ &\geq \frac{2}{\sigma}\left(\kappa_n \ln\left(\frac{\sigma}{2\epsilon}\right) - C\right)\sum_i r_i^{\text{old}}.\end{aligned}$$

With (4.3) this shows that

$$\sum_i r_i^{\text{old}} \leq d\sigma$$

if ϵ is sufficiently small. Note also from the construction of Lemma 3.1 that

$$\begin{aligned} d\sigma &\geq \sum_i r_i^{\text{old}} \\ &= \sum_j r_j^{\text{new}} \\ &\geq \sum_j |d_j^{\text{new}}| s^{\text{new}} \\ &\geq d s^{\text{new}}. \end{aligned}$$

8. Thus an amalgamation step yields a collection of balls which satisfy (4.4)–(4.6), with $s \leq \sigma$. If $s \in [\sigma/2, \sigma]$, then we are finished. If not, because the balls are pairwise disjoint, we modify them again using the expansion procedure of steps 4 and 5 above. We continue in this fashion, expanding and combining balls, until (4.7) is satisfied.

This must happen eventually, for the following reasons: each amalgamation step decreases the number of balls, and each expansion step leaves the total number of balls unchanged. Thus there can be only finitely many amalgamations. Also, as remarked above, there can only be finitely many expansions. It is therefore clear that the process must eventually terminate. By construction this can happen only when $s \in [\sigma/2, \sigma]$. □

We now complete the proof of Theorem 1.2:

Proof. We may assume that (4.1)–(4.3) hold, if necessary by extending u in a neighborhood of ∂U, as in Lemma 4.1.

By a standard approximation argument, we may assume that u is continuous.

We may thus construct balls as in Proposition 4.1. By an obvious modification of Proposition 4.1, we may prescribe that $s := \min r_i/|d_i| \in [\sigma/4, \sigma/2]$.

Let $\{x_1, \ldots, x_m\}$ be the centers of the balls which have nonzero degree, and let $\{d_1, \ldots, d_m\}$ be the corresponding degrees. In view of (4.8) it is clear that $d_i > 0$ for all i. From (4.6), (4.2), and the definition of S_E, we deduce that $\sum d_i = d$.

Finally, from (4.3), (4.4), (4.5), and (3.4) we deduce that $\sum_i (r_i/s) \leq d + 1$ for ϵ small. Since $d_i \leq r_i/s$, this implies that $r_i \leq 2d_i s \leq d_i \sigma \ \forall\ i$. Thus

$$B_i \subset B_{d_i \sigma}, \tag{4.16}$$

and so we can use familiar properties of Λ^ϵ to estimate

$$\begin{aligned} \int_{\cup_i B_{d_i\sigma}(x) \cap U} \mathcal{E}^\epsilon \, dx &\geq \int_{\cup_i B_i \cap U} \mathcal{E}^\epsilon \, dx \\ &\geq \sum_i \frac{r_i}{s} \Lambda^\epsilon(s) \\ &\geq \sum_i d_i \Lambda^\epsilon(s) \\ &\geq d\Lambda^\epsilon\left(\frac{\sigma}{4}\right) \\ &\geq d\kappa_n \ln\left(\frac{\sigma}{\epsilon}\right) - C. \end{aligned}$$

□

5. Compactness. Theorem 1.2 implies a compactness result in a straightforward way. The point is that, given a sequence of functions, we can find a subsequence along which the energy is uniformly bounded away from a collection of at most d limiting singular points.

In this section we will prove the following proposition.

PROPOSITION 5.1. *Suppose u^ϵ is a collection of functions and that for each ϵ, u^ϵ satisfies* (4.1), (4.2), *and*

$$\int_U \mathcal{E}^\epsilon dx \le d\kappa_n \ln\left(\frac{1}{\epsilon}\right) + C.$$

Then there exist points $x_1, \dots, x_m \in \bar{U}$, with $m \le d$, a subsequence $\epsilon_k \to 0$, and a function $u \in W^{1,n}_{loc}(U \setminus \{x_1, \dots, x_m\}; S^{n-1})$ such that

$$u^{\epsilon_k} \rightharpoonup u \qquad \text{weakly in } W^{1,n}_{loc}(U \setminus \{x_1, \dots, x_m\}; \mathcal{R}^n).$$

Finally, there are integers $d_i > 0$ for $i = 1, \dots, m$ such that $\sum d_i = d$ and

$$\mu^{\epsilon_k} := |\ln \epsilon_k|^{-1} \mathcal{E}^{\epsilon_k} dx \to \kappa_n \sum_{i=1}^m d_i \delta_{x_i}$$

weakly as measures.

Remark. Theorem 1.3 follows immediately from this proposition and Lemma 4.1.

Proof. 1. Fix $\sigma_1 \in (0, r/4d)$, and for every $\epsilon < \epsilon_0(\sigma)$, apply Proposition 4.1 to the function u^ϵ. This yields a collection of balls $\{B_i^\epsilon\}_{i=1}^M$. Let B_i^ϵ have center x_i^ϵ, and let $d_i^\epsilon := \deg(u^\epsilon; \partial B_i^\epsilon)$. Discard all balls for which $d_i^\epsilon = 0$. Then at most d balls remain, the balls are disjoint, and each ball satisfies

$$\int_{B_i^\epsilon} \mathcal{E}^\epsilon dx \ge d_i^\epsilon \Lambda^\epsilon \left(\frac{\sigma_1}{2}\right) \ge d_i^\epsilon \left[\kappa_n \ln\left(\frac{\sigma_1}{\epsilon}\right) - C\right] \tag{5.1}$$

by construction. Arguing as in (4.16), we see that

$$B_i^\epsilon \subset B_{2d_i^\epsilon \sigma_1}(x_i^\epsilon). \tag{5.2}$$

We may easily find positive integers d_i^1, points x_i^1, and a subsequence $\epsilon_j \to 0$ such that

$$d_i^{\epsilon_j} \to d_i^1, \qquad \sum_i d_i^1 = d,$$

and for each i,

$$x_i^{\epsilon_j} \to x_i^1$$

as $j \to \infty$.

Define $B_i^1 := B_{4d_i^1 \sigma_1}(x_i^1)$, and

$$U^1 := U \setminus (\cup_i B_i^1).$$

From (5.2) we see that $B_i^{\epsilon_j} \subset B_i^1$ for all sufficiently large j. Thus (4.3) and (5.1) imply that

$$\liminf_j \int_{B_i^1} \mathcal{E}^{\epsilon_j} dx - d_i^1 \kappa_n \ln\left(\frac{\sigma_1}{\epsilon}\right) \ge -C, \tag{5.3}$$

$$\limsup_j \int_{U^1} \mathcal{E}^{\epsilon_j}\,dx \le d\kappa_n \ln\left(\frac{1}{\sigma_1}\right) + C. \tag{5.4}$$

2. Now repeat the above process for $\sigma_2 = \frac{1}{2}\sigma_1$ to find positive integers d_i^2, points x_i^2, and a further subsequence, which we still write as ϵ_j, such that (5.3) and (5.4) hold with σ_1 replaced with σ_2, where B_i^2 and U^2 are defined as before.

Note that for any j,

$$B_j^1 \cap (\cup_i B_i^2) \neq \emptyset \qquad \text{and} \qquad B_j^2 \cap (\cup_i B_i^1) \neq \emptyset.$$

If this is not the case, then, for example,

$$B_j^2 \subset U^1,$$

which would lead to a contradiction between (5.4) and (5.1).

It follows that

$$\operatorname{dist}\left(\{x_i^1\}_i\,,\,\{x_i^2\}_i\right) \le \sigma_1 + \sigma_2, \tag{5.5}$$

where dist here denotes the Hausdorff distance between two sets,

$$\operatorname{dist}(A,B) = \max\{\max_{x\in A}\min_{y\in B}|x-y|\,,\,\max_{y\in B}\min_{x\in A}|x-y|\}.$$

3. We now repeat the same procedure for each k, with $\sigma_k = 2^{-k+1}\sigma_1$, each time passing to subsequences and finding points x_i^k and positive integers d_i^k such that $\sum_i d_i^k = d$,

$$\liminf_j \int_{B_i^k} \mathcal{E}^{\epsilon_j}\,dx - d_i^k\kappa_n \ln\left(\frac{\sigma_k}{\epsilon}\right) \ge\ -C, \tag{5.6}$$

and

$$\limsup_j \int_{U^k} \mathcal{E}^{\epsilon_j}\,dx \le d\kappa_n \ln\left(\frac{1}{\sigma_k}\right) + C \tag{5.7}$$

for balls B_i^k and U^k defined as in step 1.

By a diagonal argument we may extract a subsequence, still denoted ϵ_j, such that (5.6) and (5.7) hold for every k. Moreover it is clear from (5.5) that the sets $\{x_i^k\}_i$ must converge in the Hausdorff metric to some limiting set $\{x_1,\dots,x_m\}$ as $k\to\infty$. Any compact subset of $U\setminus\{x_1,\dots,x_m\}$ is contained in some U_k for k sufficiently large. Thus $\{u^{\epsilon_j}\}$ is weakly precompact in $W^{1,n}_{\text{loc}}(U\setminus\{x_1,\dots,x_m\};\mathcal{R}^n)$.

The other conclusion of the theorem follows similarly from (5.6) and (5.7). ∎

6. The gauge-invariant functional. The above framework can be modified to give estimates for the gauge-invariant functional $I^\epsilon_{\text{mag}}(u,A)$.

Let $\gamma := ||\nabla\times A||_{L^2(U)}$ and

$$\mathcal{F}^\epsilon_{\text{mag}}[u,A] := \frac{1}{2}|\nabla_A u|^2 + \frac{1}{4\epsilon^2}(1-|u|^2)^2,$$

so that $I^\epsilon_{\text{mag}}(u,A) = \frac{1}{2}\gamma^2 + \int_U \mathcal{F}^\epsilon_{\text{mag}}[u,A]\,dx$.

We also define

$$\lambda^\epsilon_\gamma(r;d) := \min_{m\in[0,1]}\left\{\frac{1}{C\epsilon}|1-m|^N + \frac{m^2}{r}\left[\left(\sqrt{\pi}d - \frac{r\gamma}{2}\right)^+\right]^2\right\}. \tag{6.1}$$

We will regard γ as fixed and establish lower bounds for $\mathcal{F}^\epsilon_{\text{mag}}$ in which γ appears as a parameter. These bounds have exactly the same structure as our earlier bounds, so that the covering arguments from earlier sections may by employed almost without change to find global lower bounds for $\mathcal{F}^\epsilon_{\text{mag}}$ which depend on the parameter γ. We then obtain estimates for the full energy I^ϵ_{mag} by minimizing over $\gamma > 0$.

LEMMA 6.1. *λ^ϵ_γ has the following properties. First,*

$$\lambda^\epsilon_\gamma(r;d) \geq \frac{\pi}{r}\left[\left(d - \frac{r\gamma}{2\sqrt{\pi}}\right)^+\right]^2 \left[1 - C\frac{\epsilon^\alpha}{r^\alpha}\right]. \tag{6.2}$$

Second, if $r \geq \epsilon$, $B_r \subset U$, $u \in H^1(\partial B_r, \mathcal{C})$, and $|\deg(u;\partial B_r)| = d$, then

$$\int_{\partial B_r} \mathcal{F}^\epsilon_{mag}[u,A]\, dH^1 \;\geq\; \lambda^\epsilon_\gamma(r;d). \tag{6.3}$$

Proof. 1. We omit the proof of (6.2), as it is exactly like the proof of (2.2).

2. We may assume without loss that $d > 0$. We write $u = \rho e^{i\phi}$, where ϕ is multivalued. A calculation shows that

$$|\nabla_A u|^2 = |D\rho|^2 + \rho^2|D\phi - A|^2.$$

Thus

$$\int_{\partial B_r} \mathcal{F}^\epsilon_{\text{mag}}\, dH^1 = I_1 + I_2,$$

for

$$I_1 := \int_{\partial B_r} \frac{1}{2}|D\rho|^2 + \frac{1}{4\epsilon^2}(1-\rho^2)^2,$$

$$I_2 := \int_{\partial B_r} \frac{1}{2}\rho^2|D\phi - A|^2.$$

Define $m := \inf_{\partial B_r} \rho$. Exactly as in Lemma 2.3 we see that

$$I_1 \geq \frac{1}{C\epsilon}|1-m|^N$$

for some $C, N > 0$. It is at this stage that we use the assumption $r \geq \epsilon$.

3. For $x := (x_1, x_2) \in \partial B_r$, let $\tau(x) := \frac{1}{|x|}(-x_2, x_1)$ denote the oriented tangent at x.

Because u has degree d,

$$\int_{\partial B_r} D\phi \cdot \tau\, dH^1 = 2\pi d.$$

Also, by Stokes' theorem,

$$\begin{aligned}\int_{\partial B_r} A \cdot \tau\, dH^1 &= \int_{B_r} \nabla \times A \\ &\leq (\pi r^2)^{1/2}\left(\int_{B_r} |\nabla \times A|^2\right)^{1/2} \\ &\leq \sqrt{\pi} r\gamma.\end{aligned}$$

Thus

$$2\pi d - \sqrt{\pi} r\gamma \;\le\; \int_{\partial B_r} (D\phi - A)\cdot\tau,$$

and therefore

$$(2\pi d - \sqrt{\pi} r\gamma)^+ \le \sqrt{2\pi} r \left(\int_{\partial B_r} |D\phi - A|^2 \right)^{1/2}.$$

Here $s^+ := \max\{s, 0\}$ for $s \in \mathcal{R}$, as usual. Rearranging and using the definition of m, we obtain

$$I_2 \ge \frac{m^2}{r} \left[\left(\sqrt{\pi} d - \frac{r\gamma}{2} \right)^+ \right]^2$$

With step 2 this immediately gives the conclusion of the lemma. □

Next we define

$$\Lambda^\epsilon_\gamma(r) := \int_0^r \lambda^\epsilon_\gamma(s;1) \wedge \frac{c_0}{\epsilon} ds \tag{6.4}$$

for some appropriately small c_0. We continue to assume that $\gamma = ||\nabla \times A||_{L^2}$ is some known, finite number.

Λ^ϵ_γ has the following properties.

PROPOSITION 6.1. *$\Lambda^\epsilon_\gamma(\cdot)$ is increasing; moreover,*

$$\Lambda^\epsilon_\gamma(r+s) \le \Lambda^\epsilon_\gamma(r) + \Lambda^\epsilon_\gamma(s) \qquad \forall\, r, s \ge 0; \tag{6.5}$$

$$s \mapsto \frac{1}{s}\Lambda^\epsilon_\gamma(s) \qquad \textit{is nonincreasing}; \qquad \textit{and} \tag{6.6}$$

$$\Lambda^\epsilon_\gamma(r) \;\ge\; \pi \log\left(\frac{1}{\epsilon}\right) + \log(r \wedge \gamma^{-1}) - C. \tag{6.7}$$

Proof. The lower estimate (6.7) follows by integrating the lower bound (6.2) for λ^ϵ_γ. The other claims are proven by exactly the arguments used in the proof of Proposition 3.1. □

The next proposition follows from (6.3) by exactly the arguments of Proposition 3.2.

PROPOSITION 6.2. *Let $u \in H^1(U;\mathcal{C})$ be continuous. If $|\mathrm{dg}\,(u;\partial B_\rho)| = d > 0$ for all $\rho \in [r_0, r_1]$ and $\epsilon \le r_0 \le r_1$, then*

$$\int_{B_{r_1}\setminus B_{r_0}} \mathcal{F}^\epsilon_{mag}[u,A]\,dx \ge d\left[\Lambda^\epsilon_\gamma\left(\frac{r_1}{d}\right) - \Lambda^\epsilon_\gamma\left(\frac{r_0}{d}\right)\right]. \tag{6.8}$$

Also,

$$\int_{B_{r_1}\setminus B_{r_0}} \mathcal{F}^\epsilon{}_{mag}\,[u,A]\,dx \ge \Lambda^\epsilon_\gamma(r_1) - \Lambda^\epsilon_\gamma(r_0). \tag{6.9}$$

□

The final property of Λ that we will need is the following. For this we assume that $u \in C \cap H^1(U;\mathcal{C})$, and we define S_E as in (1.10).

PROPOSITION 6.3. *Suppose that $u \in C \cap H^{1,2}(U;\mathcal{C})$ and that $A \in H^1(U;\mathcal{R}^2)$, with $\gamma := ||\nabla \times A||_{L^2} \leq C\epsilon^{-1/2}$. If $S_E \subset\subset U$, then there is a collection of closed, pairwise disjoint balls $\{B_i\}_{i=1}^k$ such that*

$$S_E \subset \cup_{i=1}^k B_i,$$

$$\int_{B_i \cap U} \mathcal{F}^\epsilon_{mag} dx \geq \Lambda^\epsilon_\gamma(r_i),$$

$$r_i \geq \epsilon \qquad \forall\, i,$$

$$B_i \cap S_E \neq \emptyset \qquad \text{for each } i.$$

Proof. The proof follows exactly that of Proposition 3.3, which uses only Lemmas 2.5, 3.1, and 3.2. In this context, the first two of these are still valid, and in place of Lemma 3.2, we have Lemma 6.2 below, so the desired result is a consequence of our earlier arguments. □

LEMMA 6.2. *Let S_i be a connected component of S_E, and assume that $S_i \subset\subset U$. Assume also that $\gamma \leq C\epsilon^{-1/2}$. Then there exists some constant C such that, for all sufficiently small ϵ,*

$$\int_{S_i} \mathcal{F}^\epsilon[u,A]dx \geq C^{-1}.$$

Proof. 1. We may assume by an approximation argument that u is C^∞. As above we write $u := \rho e^{i\phi}$.

Let $f := \frac{1}{2}\rho^2$. We will use the coarea formula,

$$\int_{S_i} g\, Jf\, dx = \int_0^\infty \left(\int_{S_i \cap f^{-1}(s)} g dH^{n-1} \right) ds, \tag{6.10}$$

where Jf is the Jacobian of f, which in this case is given by

$$Jf = |Df| = \rho |D\rho|.$$

Thus

$$\begin{aligned} |\nabla_A u|^2 &= |D\rho|^2 + \rho^2 |D\phi - A|^2 \\ &= Jf \left(\frac{|D\rho|}{\rho} + \frac{\rho}{|D\rho|} |D\phi - A|^2 \right) \\ &\geq Jf |D\phi - A|. \end{aligned}$$

The coarea formula thus gives

$$\begin{aligned} \int_{S_i} |\nabla_A u|^2 &\geq \int_0^\infty \left(\int_{f^{-1}(s)\cap S_i} |D\phi - A| dH^1 \right) ds \\ &\geq \int_0^\infty \left| \int_{f^{-1}(s)\cap S_i} (D\phi - A)\cdot \tau dH^1 \right| ds. \end{aligned}$$

2. For $s \in [0,1]$, $f^{-1}(s) \cap S_i$ is nonempty, since u is continuous and must have a zero in each component of S_E. Moreover, for Lebesgue (almost everywhere) such s, $f^{-1}(s) \cap S_i$ is a smooth compact manifold, by Sard's theorem. Fix some s for which this is the case. The definition of degree (1.6) easily implies that

$$\deg(u^\epsilon; f^{-1}(s) \cap S_i) = \deg(u^\epsilon; \partial S_i) := d.$$

It follows that

$$\int_{f^{-1}(s)\cap S_i} D\phi \cdot \tau = d.$$

Also,

$$\begin{aligned}\int_{f^{-1}(s)\cap S_i} A \cdot \tau \, dH^1 &= \int_{\{\frac{1}{2}\rho^2 \le s\}\cap S_i} \nabla \times A \, dx \\ &\le |S_i|^{1/2} \left(\int_{S_i} |\nabla \times A|^2 \, dx \right)^{1/2} \\ &\le |S_i|^{1/2} \gamma.\end{aligned}$$

Putting these together, we find that

$$\int_{f^{-1}(s)\cap S_i} (D\phi - A) \cdot \tau dH^1 \ge d - |S_i|^{1/2}\gamma.$$

3. Combining steps 1 and 2 and using the fact that $(1-|u|^2)^2 \ge C^{-1}$ on S_i, we obtain

$$\int_{S_i} \mathcal{F}_{\text{mag}} dx \ge d - |S_i|^{1/2}\gamma + |S_i| \frac{1}{C\epsilon^2}.$$

Since we have assumed that $\gamma \le C\epsilon^{-1/2}$, the conclusion follows by calculus. □

Using Propositions 6.1, 6.2, and 6.3, we can repeat the proofs of Proposition 4.1 and Theorem 1.2 with only minor modifications. We indicate how this is done.

PROPOSITION 6.4. *Assume that u is continuous and satisfies* (4.1)–(4.2) *and that*

$$\int \mathcal{E}^\epsilon_{mag}[u, A] \, dx \le \pi d \ln\left(\frac{1}{\epsilon}\right) + C. \tag{6.11}$$

Then

$$\int_U |\nabla \times A|^2 \, dx \le C. \tag{6.12}$$

Moreover, given any $\sigma \in (0, \frac{r}{4d})$, there exists some $\epsilon_0(\sigma) > 0$ such that, for each $\epsilon \in (0, \epsilon_0)$, we can find a collection of closed balls $\{B_i\}_{i=1}^M$ of radius r_i and degree d_i such that

the interiors of the balls are pairwise disjoint, (6.13)

$$\int_{B_i \cap U} \mathcal{F}^\epsilon_{mag}[u, A] \, dx \ge \frac{r_i}{s} \Lambda^\epsilon(s), \qquad \text{for } s := \min_j r_j/|d_j|, \tag{6.14}$$

$$S_E \subset\subset \bigcup_{i=1}^{M} B_i, \qquad \text{and } B_i \cap S_E \neq \emptyset \text{ for every } i. \tag{6.15}$$

Moreover,

$$s \in \left[\frac{\sigma}{2}, \sigma\right], \tag{6.16}$$

$$d_i \geq 0 \qquad \forall i. \tag{6.17}$$

Proof. 1. Fix some $\sigma \in (0, \frac{r}{4d})$. Following step 1 of the proof of Proposition 4.1, we show that it suffices to find a collection of balls satisfying (6.13)–(6.16).

Suppose that we have found such a collection; we need to show that (6.12) and (6.17) necessarily hold. Following (4.10),

$$\begin{aligned}\int_U \mathcal{E}^\epsilon_{\text{mag}}\, dx &\geq \frac{1}{2}\gamma^2 + \sum_i \int_{B_i} \mathcal{F}^\epsilon_{\text{mag}}\, dx \\ &\geq \frac{1}{2}\gamma^2 + \Lambda^\epsilon_\gamma\left(\frac{\sigma}{2}\right)\sum |d_i| \\ &\geq \frac{1}{2}\gamma^2 + \pi\left[\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{\sigma}{2}\wedge\frac{1}{\gamma}\right)\right]\sum |d_i|. \qquad (6.18)\end{aligned}$$

If $d_i < 0$ for some i, then $\sum |d_i| \geq d+2$. If this holds, then (6.18) and (6.11) yield

$$\pi d \ln\left(\frac{1}{\epsilon}\right) + C \;\geq\; \frac{1}{2}\gamma^2 + (d+2)\pi\left[\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{\sigma}{2}\wedge\frac{1}{\gamma}\right)\right].$$

By minimizing the right-hand side over $\gamma > 0$, we obtain a contradiction if ϵ is sufficiently small. This proves that $d_i \geq 0\ \forall\ i$.

Once we know that $\sum |d_i| = d$, by comparing (6.18) and (6.11) we obtain

$$\frac{1}{2}\gamma^2 + d\pi \ln\left(\frac{\sigma}{2}\wedge\frac{1}{\gamma}\right) \;\leq\; C.$$

This clearly implies that $\gamma \leq C$, which is (6.12).

2. As in step 2 of the proof of Proposition 4.1, we next show that if a collection of balls satisfies (6.13)–(6.15), and if $s \leq \sigma$, then $B_i \cap \partial U = \emptyset$ for every ball in the collection.

Suppose, toward a contradiction, that $B_i \cap \partial U \neq \emptyset$ for some i. By following exactly the argument that leads to (4.11), we see that

$$\begin{aligned}\int_U \mathcal{E}^\epsilon_{\text{mag}}\, dx &\geq \frac{1}{2}\gamma^2 + \int_{B_i} \mathcal{F}^\epsilon_{\text{mag}}\, dx \\ &\geq \frac{1}{2}\gamma^2 + \frac{r_i}{\sigma}\Lambda^\epsilon_\gamma(\sigma) \\ &\geq \frac{1}{2}\gamma^2 + 2d\pi\left[\ln\left(\frac{1}{\epsilon}\right) + \ln\left(\frac{\sigma}{2}\wedge\frac{1}{\gamma}\right)\right].\end{aligned}$$

This leads to a contradiction with (6.11) for sufficiently small ϵ, again by minimizing over $\gamma \geq 0$.

3. Steps 3 through 6 and step 8 of the proof of Proposition 4.1 may be repeated without change. Step 7 requires some small modifications, which can be carried out along the lines already indicated above. ∎

Acknowledgment. I am indebted to Mete Soner for pointing out the result of Lemma 2.3.

REFERENCES

[1] L. Almeida and F. Bethuel, *Topological methods for the Ginzburg-Landau equations*, J. Math. Pures Appl., 77 (1998), pp. 1–49.

[2] F. Bethuel, H. Brezis, and F. Hélein, *Ginzburg-Landau Vortices, Progress in Nonlinear Differential Equations and their Applications* 13, Birkhäuser Boston, Cambridge, MA, 1994.

[3] F. Bethuel and T. Rivière, *Vortices for a variational problem related to superconductivity*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 12 (1995), pp. 243–303.

[4] H. Brezis, F. Merle, and T. Rivière, *Quantization effects for* $-\Delta u = u(1 - |u|^2)$ *in* $\mathbf{R}^2$, Arch. Rational Mech. Anal., 126 (1994), pp. 35–58.

[5] H. Brezis and L. Nirenberg, *Degree theory and BMO.* I. *Compact manifolds without boundaries*, Selecta Math. (N.S.), 1 (1995), pp. 197–263.

[6] Z. Han and Y. Li, *Degenerate elliptic systems and applications to Ginzburg–Landau type equations.* I, Calc. Var. Partial Differential Equations, 4 (1996), pp. 171–202; *Erratum*, Calc. Var. Partial Differential Equations, 4, (1996) p. 497.

[7] M.-C. Hong, *Asymptotic behavior for minimizers of a Ginzburg-Landau-type functional in higher dimensions associated with n-harmonic maps*, Adv. Differential Equations, 1 (1996), pp. 611–634.

[8] F. Lin, *Some dynamical properties of Ginzburg-Landau vortices*, Comm. Pure. Appl. Math, 49 (1996), pp. 323–359.

[9] F. Lin, *Vortex Dynamics for the Nonlinear Wave Equation*, preprint, 1997.

[10] E. Sandier, *Lower Bounds for the Energy of Unit Vector Fields and Applications*, preprint, 1997.

[11] M. Struwe, *On the asymptotic behavior of minimizers of the Ginzburg-Landau model in* 2 *dimensions*, Differential Integral Equations, 7 (1994), pp. 1613–1624.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 747–755

STRUCTURAL STABILITY FOR THE EULER METHOD*

MING-CHIA LI†

Abstract. In this paper, we show that if a flow φ^t has a hyperbolic chain recurrent set either without fixed points or with only fixed points, and satisfies the strong transversality condition, then φ^t is structurally stable with respect to numerical methods, including the Euler method, which was not done in [B. M. Garay, *Numer. Math.*, 72 (1996), pp. 449–479], [M.-C. Li, *J. Differential Equations*, 141 (1997), pp. 1–12], and [M.-C. Li, *SIAM J. Math. Anal.*, 28 (1997), pp. 381–388]. The proof is an application of the invariant manifold techniques developed by Hirsch, Pugh, and Shub [M. Hirsh, C. Pugh, and M. Shub, Invariant Manifolds, *Lecture Notes in Math.* 583, *Springer-Verlag, New York,* 1977] and Robinson [C. Robinson, *J. Differential Equations*, 22 (1976), pp. 28–73]. The result is an extension of our previous work [M.-C. Li, *Proc. Amer. Math. Soc.*, 127 (1999), pp. 289–295].

Key words. structural stability, hyperbolicity, chain recurrent set, Euler method, numerical method

AMS subject classifications. 58F09, 58F10, 34D30, 65L06, 65L20

PII. S0036141097322558

1. Introduction.

How much confidence can we have in computer simulation? For example, suppose we use a numerical method to approximate the general solution of a differential equation. Does the numerical approximation exhibit the same global geometry and dynamics as the exact solution? There have been theorems showing that the answer to this question is yes for several cases. Before going into these theorems, we give an introduction on how the global qualitative theory in dynamical systems applies to numerical analysis.

Numerical methods and dynamical systems are deeply connected. As is well known, general solutions of differential equations are considered continuous-time dynamical systems. On the other hand, numerical methods frequently involve iterative processes, and so they are discrete-time dynamical systems. Suppose we use a numerical method to approximate the general solution of a differential equation. The numerical method can be interpreted as a small perturbation of the solution flow. The natural question is whether the numerical approximation accurately reflects the dynamics of the original flow. In the terminology of global stabilities, this is the same as asking whether certain flows are structurally stable with respect to numerical methods.

There are many classical stability theorems in dynamical systems. Local stability results include the Hartman–Grobman theorem, the stable manifold theorem, and the normally hyperbolic stability theorem. For global results, we have the Ω-stability theorem and the structural stability theorem. The structural stability theorem states that a continuous/discrete-time dynamical system preserves its global geometric structure under small C^1 perturbations. Here, we have to point out that results about structural stability cannot be applied directly to our situation: a flow is approximated by a numerical method which is a discrete-time system, not a continuous system. In order to show structural stability of flows for numerical methods, we need to make

*Received by the editors June 2, 1997; accepted for publication (in revised form) July 23, 1998; published electronically May 7, 1999.
http://www.siam.org/journals/sima/30-4/32255.html
†Department of Mathematics, National Changhua University of Education, Changhua 500, Taiwan (mcli@math.ncue.edu.tw).

explicit estimations by going through the mathematical techniques in the proof of the structural stability theorem.

Historically, the structural stability theorem was attacked for several special cases before the general proof was given. First, the case of Anosov systems when the whole manifold has a hyperbolic structure was proved by Anosov [1] and Moser [21]. Next, the case of Morse–Smale systems when the nonwandering set is a finite number of hyperbolic closed orbits was proved by Palis and Smale [22]. Finally, the case of Axiom A systems when the nonwandering set is the union of finitely many basic sets was proved by Robbin [25] for C^2 diffeomorphisms, by Robinson [26] for C^2 vector fields, and again by Robinson for C^1 diffeomorphisms [28] and C^1 flows [27].

Suppose that we have a flow φ^t on a compact manifold M which is approximated by a numerical method N in the following sense: the numerical method of stepsize h and order p, N^h, is $O(h^{p+1})$-close to the time-h map of the flow, φ^h. Under certain differentiability conditions on φ and N, one can show that N^h is $O(h^p)$-close to φ^h in the C^1 topology (see Lemma 1). We want to match the trajectories of φ^t with the orbits of N^h by constructing a homeomorphism H_h, depending on h, to conjugate N^h to φ. In terms of dynamical systems, H_h is usually called a *topological conjugacy*, if there is no reparametrization on φ^t, or a *topological equivalence* otherwise. If such a homeomorphism H_h exists, we say that the flow φ^t is *structurally stable* with respect to the numerical method N.

Global stability under numerical approximation is analogous to the structural stability theorem. Two analytic approaches were used by Robinson in [28] to prove the structural stability theorem. One involves solving a functional equation by means of the implicit function theorem. The other involves using compatible families of stable and unstable disks by the method of graph transforms.

Certain flows are structurally stable with respect to numerical methods. The case of Morse–Smale gradient-like flows, i.e., Morse–Smale flows without closed orbits, was proved by Garay [12]. See also [18] for a slightly better result. The case of Axiom A flows with the strong transversality condition was shown in [17]. The proof of these theorems is essentially based on the first approach in [28]: solve a functional equation to construct a conjugacy H_h and use the d_φ-Lipschitz metric, due to Robbin [25], to prove H_h is one to one. These theorems work well for numerical methods of order $p \geq 2$. If we consider the Euler method of order $p = 1$, then we get that H_h is a semiconjugacy but cannot prove it is one to one.

The challenge for the Euler method led us to consider the second approach in [28] using the graph transforms. In [19], we assumed φ^t has a hyperbolic attractor $\mathcal{A}_\varphi$. Relying on the abstract invariant manifold theorems in [28] and the idea of laminations given by Hirsch, Pugh, and Shub [16], we showed that the restriction of the flow φ^t to the basin of attraction, $B(\mathcal{A}_\varphi)$, is structurally stable with respect to numerical methods of order $p \geq 1$. Earlier, Garay [13] showed that the restriction of φ^t to $B(\mathcal{A}_\varphi)\backslash\mathcal{A}_\varphi$, $\varphi^t|B(\mathcal{A}_\varphi)\backslash\mathcal{A}_\varphi$, is structurally stable under discretizations.

In this paper, we extend our previous result [19] to the case when the flow φ^t has a hyperbolic chain recurrent set but no fixed points. By using compatible families of stable and unstable disks due to Robinson [28], we are able to construct a topological equivalence and prove that φ^t is structurally stable with respect to numerical methods, including the Euler method. The proof is even simpler for the case when the chain recurrent set has only fixed points.

It is not clear whether we can use this method to prove that φ^t with hyperbolic fixed and periodic points simultaneously is structurally stable with respect to the Eu-

ler method. Proving the structural stability theorem for flows, Robinson [28] put a reparametrization on the perturbed flow and successfully made a transition from no reparametrization near fixed points to a reparametrization away from fixed points. Here, we cannot parametrize the numerical method since it is discrete. Recently, Pilyugin [23] showed a shadowing result for structurally stable flows by dealing with different hyperbolic structures near fixed points and near periodic points.

Although not emphasized here, the persistence of local properties under approximation has been widely studied. See [4], [10] for hyperbolic equilibria, [3], [6], [8], [9], [24] for hyperbolic closed orbits, and [7], [11], [14], [20], [24] for normally hyperbolic invariant manifolds. Most of such results are discussed in a survey by Stuart [36] (see also [37]).

2. Statement of theorems. First, we introduce notations and basic definitions.

Let M be a smooth complete Riemannian manifold with a distance d arising from the Riemannian metric and $\mathrm{Diff}(M)$ be the set of diffeomorphisms on M with the strong topology and distance d_{C^1}. A *flow* is a map $\varphi : \mathbb{R} \times M \to M$ that satisfies the group property $\varphi(s, \varphi(t, x)) = \varphi(s + t, x)$. We write $\varphi(t, x) = \varphi^t(x)$.

A compact invariant set Λ for a C^1 flow φ^t on M is *hyperbolic* if the restriction of the tangent bundle TM of M to Λ splits into three continuous subbundles, $TM|\Lambda = \mathbb{E}^u \oplus \mathbb{E}^s \oplus Span(X)$, invariant under the derivative of φ^t, $D\varphi^t$, such that $D\varphi^t$ expands $\mathbb{E}^u$ and $D\varphi^t$ contracts $\mathbb{E}^s$ for all $t \geq t_0$, where X is the vector field induced by the flow φ^t. The *stable manifold* of $x \in M$ is the set of all points in M whose forward orbits under φ^t tend to x. The *unstable manifold* of $x \in M$ is the set of all points in M whose backward orbits under φ^t tend to x. The flow φ^t satisfies the *strong transversality condition* if all the stable manifolds of orbits of φ^t transversally intersect the unstable manifolds of orbits of φ^t. See [30] for a more complete explanation of these definitions.

Chain recurrent sets for flows are defined as follows.

Definition 1. *A pair of sets $(\mathcal{A}, \mathcal{A}^*)$ is called the* attractor-repeller pair *for the flow φ^t if there exists a neighborhood U of $\mathcal{A}$ such that $\varphi^T(\mathrm{cl}(U)) \subset \mathrm{int}(U)$ for some $T > 0$, $\mathcal{A} = \cap_{t \geq 0} \varphi^t(U)$, and $\mathcal{A}^* = \bigcap_{t \leq 0} \varphi^t(M \backslash U)$. The* chain recurrent set $\Re_\varphi$ *for φ^t is the set $\Re_\varphi = \cap_{i \in I}(\mathcal{A}_i \cup \mathcal{A}^*{}_i)$, the intersection of all attractor-repeller pairs.*

One can define the chain recurrent set to be the points x such that there is a periodic ϵ-chain through x for all $\epsilon > 0$. These two definitions are equivalent; see [30] for the proof. In our previous results on structural stability for numerical methods, we assumed hyperbolicity on the nonwandering set Ω_φ in [17] and on an attractor $\mathcal{A}_\varphi$ in [19]. In the present paper, the hyperbolicity will be assumed on the chain recurrent set $\Re_\varphi \supset \Omega_\varphi \supset \mathcal{A}_\varphi$.

We use a general definition of numerical methods on manifolds.

Definition 2. *Let φ^t be a flow on M. For $p \geq 1$, a C^{p+1} function $N : \mathbb{R} \times M \to M$ is called a* numerical method of order p *for φ^t if there are positive constants K and h_0 such that $d(\varphi^h(x), N^h(x)) \leq Kh^{p+1}$, for all $h \in [0, h_0]$ and $x \in M$. Here h stands for a* stepsize *of N. We denote the ith iterate of $N^h(x)$ by $(N^h)^i(x)$.*

The definition is related to standard numerical analysis. For example, the explicit Euler method in Euclidean spaces corresponds to $N^h(x) = x + h\varphi^h(x)$ and is of order $p = 1$. Both the improved Euler method and the three term Taylor series method are of order 2. The usual Runge–Kutta methods are of order $p \geq 4$. See [2] and [5]. The pth-order multiderivative method is another example satisfying the above conditions (see [15]).

The distance between φ^h and N^h in the C^1-topology is estimated in the following lemma.

LEMMA 1 (see [11]). *Let N be a numerical method of order p for a C^{p+1} flow φ^t on a compact manifold M. Then there is a positive constant K_1 such that $d_{C^1}(\varphi^h, N^h) \le K_1 h^p$ for all sufficiently small h. Moreover, given $T > 0$, there is a positive constant K_2 such that $d_{C^1}(\varphi^T, (N^{\frac{T}{n}})^n) \le K_2 n^{1-p}$ for all large $n \in \mathbb{N}$.*

For convenience, we give the theorem for the general numerical methods first and state those for the Euler method later. The proof of the following theorem is postponed until section 4.

THEOREM 1. *Let M be a compact manifold and φ^t be a C^1 flow on M such that* (i) *φ^t has no fixed point,* (ii) *the chain recurrent set $\Re_\varphi$ is hyperbolic, and* (iii) *φ^t satisfies the strong transversality condition. Let $p \ge 1$, $T > 0$, and N be a numerical method of order p for φ^t satisfying $d_{C^1}(\varphi^T, (N^{\frac{T}{n}})^n) \le K n^{-1}$ for all large $n \in \mathbb{N}$, where K is a positive constant. Then for all sufficiently large n, there is a homeomorphism H_n on M and a continuous function $\tau_n : M \to \mathbb{R}$ such that for all $x \in M$, $H_n \circ \varphi^{\tau_n(x)}(x) = (N^{\frac{T}{n}})^n \circ H_n(x)$ and $d(H_n(x), x) \to 0$ as $n \to \infty$.*

The applicability of Theorem 1 to the Euler method follows from the following result of Shub [31].

LEMMA 2. *Let X be a C^2 bounded vector field on M, φ^t be the flow of the differential equation $\dot{x} = X(x)$, and E be the Euler method for φ^t. Then for all sufficiently small h, there is a positive constant K_1 such that $d_{C^1}(\varphi^h, E^h) \le K_1 h^2$. Moreover, given $T > 0$, there is a positive constant K_2 such that $d_{C^1}(\varphi^T, (E^{\frac{T}{n}})^n) \le K_2 n^{-1}$ for all large $n \in \mathbb{N}$.*

Now we can use Theorem 1 to get the following result for the Euler method.

THEOREM 2. *Let X be a C^2 vector field on a compact manifold M without zeros such that the differential equation $\dot{x} = X(x)$ induces a flow φ^t such that* (i) *the chain recurrent set $\Re_\varphi$ is hyperbolic and* (ii) *φ^t satisfies the strong transversality condition. Let E^h be the Euler method of stepsize h for φ^t and let $T > 0$ be given. Then for all sufficiently large n, there is a homeomorphism H_n on M and a continuous function $\tau_n : M \to \mathbb{R}$ such that for all $x \in M$, $H_n \circ \varphi^{\tau_n(x)}(x) = (E^{\frac{T}{n}})^n \circ H_n(x)$ and $d(H_n(x), x) \to 0$ as $n \to \infty$.*

We can also consider Morse–Smale gradient-like systems for the Euler method which was not done in [12] and [18].

THEOREM 3. *Let X be a C^2 vector field on a compact manifold M such that the differential equation $\dot{x} = X(x)$ induces a flow φ^t satisfying* (i) *the chain recurrent set $\Re_\varphi$ is the union of a finite number of hyperbolic fixed points and* (ii) *the stable and unstable manifolds of fixed points meet transversally. Let E^h be the Euler method of stepsize h for φ^t and $T > 0$ be given. Then for all sufficiently large n, there is a homeomorphism H_n on M and a continuous function $\tau_n : M \to \mathbb{R}$ such that for all $x \in M$, $H_n \circ \varphi^T(x) = (E^{\frac{T}{n}})^n \circ H_n(x)$ and $d(H_n(x), x) \to 0$ as $n \to \infty$.*

Note that there is no reparametrization needed for the Morse-Smale gradient-like flows. This fits well with the result in [18]. For the proof of Theorem 3, see the remark at the end of section 4.

All theorems above are still true for the case when M has boundary under the additional assumption that φ^t flows into M along the boundary. These can be done by making an adaptation of the proof in [29] (see also [18]).

3. Application. This application to the zero-finding problem is due to Shub [31].

Given a nonconstant complex polynomial $p : \mathbb{C} \to \mathbb{C}$, we then can ask to find the zeros of p, i.e., those z in $\mathbb{C}$ for which $p(z) = 0$. There are two usual approaches to turn this problem into an algorithm. One is to define the Newton differential equation

$\dot{z} = -\frac{p(z)}{p'(z)}$, where p' is the (complex) derivative of p. Then consider the Euler approximation to the solution of the Newton differential equation, that is, for h a positive real, $E_N^h(z) = z - h\frac{p(z)}{p'(z)}$. When $h = 1$, this is the Newton method. The other is to consider the Euler approximation to the gradient differential equation $\dot{z} = -\frac{1}{2}\text{grad}|p|^2$.

There is an intimately close relationship between the Newton differential equation $\dot{z} = -\frac{p(z)}{p'(z)}$ and the gradient differential equation $\dot{z} = -\frac{1}{2}\text{grad}|p|^2$. It was explained by Smale [35].

LEMMA 3. *$-\frac{1}{2}\text{grad}|p|^2 = \rho(z)(-\frac{p(z)}{p'(z)})$, where $\rho : \mathbb{C} \to \mathbb{R}$ is the positive scalar function $\rho(z) = p'(z)\overline{p'(z)}$ and $\overline{p'(z)}$ denotes the complex conjugate of $p'(z)$.*

Proof. Let $p = u + iv$ on $\mathbb{C}$. Then $|p|^2 = u^2 + v^2$. Represent $\text{grad}f = f_x + if_y$, where f_x and f_y are the partial derivatives of $f(x, y)$ with respect to x and y. We have that $\text{grad}|p|^2 = (2uu_x + 2vv_x) + i(2uu_y + 2vv_y)$. On the other hand, differentiating along the real axis, we get $p' = u_x + iv_x$ and so $\overline{p'}p = (u_x - iv_x)(u + iv) = (uu_x + vv_x) + i(-uv_x + vu_x)$.

Because p is a polynomial, it satisfies the Cauchy–Riemann equations, $u_x = v_x$ and $u_y = -v_y$. Therefore, $-\frac{1}{2}\text{grad}|p|^2 = -\overline{p'}p$ and we are done. □

This lemma says that the Newton and gradient differential equations have the same solution curves. Therefore, up to changes of the stepsize ρ, the Euler schemes of the Newton and gradient differential equations are the same.

Let r be large enough so that the disk $D_r = \{z \in \mathbb{C}; |z| \le r\}$ contains all the zeros of p. The following proposition is given by Shub and Smale [32].

PROPOSITION 1. *The trajectories of the gradient differential equation point inward transversally to the boundary of D_r.*

If all zeros of p and p' are simple, then $|p(z)|^2$ is a Morse function and $-\frac{1}{2}\text{grad}|p|^2$ is transversal to the boundary of the disk. Therefore, there is a C^1 approximation Y with $Y = -\frac{1}{2}\text{grad}|p|^2$ in a neighbor of the zeros; see [33]. We say $-\frac{1}{2}\text{grad}|p|^2$ is *generically* a Morse–Smale vector field. This fact, together with Theorem 3 for $M = D_r$, which is a manifold with a boundary, gives us an explanation of why the Newton method is convergent near the zeros of the polynomial p.

4. Proof of Theorem 1. We divide the proof into several steps.

Step 1 (preliminary setup). Because $\Re_\varphi$ is hyperbolic, the tangent bundle of M along $\Re_\varphi$ splits as the sum of three bundles $TM|\Re_\varphi = \mathbb{E}^u \oplus \mathbb{E}^s \oplus Span(X)$, where $X(x) = \frac{d\varphi^t}{dt}(x)|_{t=0}$ is the tangent vector field for φ^t. Let V_0 be a small neighborhood of $\Re_\varphi$. We want the normal bundle η of φ^t to be smooth. It is no loss of generality to make a convenient choice of η : Let η^u and η^s be smooth subbundles of $TM|V_0$ approximating $\mathbb{E}^u$ and $\mathbb{E}^s$ so that $TM|V_0 = \eta^u \oplus \eta^s \oplus Span(X)$, and choose $\eta = \eta^u \oplus \eta^s$. For $\delta = u, s$, let $\eta^\delta(r) = \{v \in \eta^\delta : |v| \le r\}$ be the r disk bundles and $\eta(r) = \eta^u(r) \oplus \eta^s(r)$.

To get an idea of the space of sections, we need to define a section's slope. If $\sigma : \eta^u(r) \to \eta$ is a section, then the *slope* of σ at $v_x \in \eta^u(r)$ is

$$\limsup_{v_y \to v_x} \frac{|s(v_x) - s(v_y)|_s}{d_u(v_x, v_y)},$$

where $\sigma(v_x) = (v_x, s(v_x)) \in \eta^u \times \eta^s$, $|\cdot|_s$ is the norm on η^s, and d_u is the metric on η^u. Let $\Sigma(1, r) = \{\text{section } \sigma : \eta^u(r) \to \eta(r) \text{ such that } slope(\sigma) \le 1\}$. Putting the C^0 sup norm on $\Sigma(1, r)$ makes it a complete metric space as usual. Let $\pi^u : \eta \to \eta^u$ be a projection along η^s and $\pi^s : \eta \to \eta^s$ be a projection along η^u.

In order to prove the conjugacy is one to one, we will need the d_φ metrics on M and TM, due to Robbin [25]. First we define the d_φ topology on M by $d_\varphi(x,y) = \sup\{d(\varphi^t(x),\varphi^t(y)) : t \in \mathbb{R}\}$. Then we isometrically embed M in $\mathbb{R}^m$ for some Euclidean space. The embedding trivializes $\eta^u \subset TM \subset M \times \mathbb{R}^m$ and $\eta^s \subset TM \subset M \times \mathbb{R}^m$. Put the d_φ metric on TM by $d_\varphi(v_x, w_y) = \max\{d_\varphi(x,y), |\pi^u v - \pi^u w|, |\pi^s v - \pi^s w|\}$, where we can subtract $\pi^\sigma v$ and $\pi^\sigma w$ since they both are in $\mathbb{R}^m$.

Step 2 (definition of bundle map). We use the concept of laminations in [16] to define a bundle map F on η. For f near φ^T in the C^1 topology, let $\Theta(\tau, v_x, f) = \exp_y^{-1} \circ f \circ \exp v_x$ where $y = \varphi^\tau(x)$. There is a neighborhood $V_1 \subset V_0$ of $\Re_\varphi$, a constant $r_1 > 0$, a neighborhood $\mathcal{U}$ of φ^T in $\mathrm{Diff}(M)$, and a continuous function $\tau : \eta(r_1)|_{V_1} \times \mathcal{U} \to \mathbb{R}$ such that for all $x \in V_1$, $v_x \in \eta_x(r_1)$, and $f \in \mathcal{U}$,

$$\Theta(\tau(v_x, f), v_x, f) \in \eta_{\varphi^{\tau(v_x,f)}(x)}(r_1).$$

Here τ stands for a reparametrization of φ^t. See [27] and also page 95 of [16]. Define a bundle map F by

$$F(v_x) \equiv \Theta(\tau(v_x, \varphi^T), v_x, \varphi^T) = \exp^{-1}_{\varphi^\tau(x)} \circ \varphi^T \circ \exp v_x.$$

Then F is a C^1 bundle map on $\eta(r_1)$.

Step 3 (existence of compatible families of unstable disks). As in [34] (see also [30]), the chain recurrent set $\Re_\varphi$ satisfying the hyperbolicity condition has a *spectral decomposition* $\Re_\varphi = \Lambda_1 \cup \cdots \cup \Lambda_m$, where the Λ_i's are pairwise disjoint and each Λ_i is closed, invariant, and topologically transitive. Since φ^t satisfies the strong transversality condition, there is a partial ordering among these sets defined by $\Lambda_i \le \Lambda_j$ if and only if $W^u(\Lambda_i) \cap W^s(\Lambda_j) \neq \emptyset$. We can extend this partial ordering to a total ordering and reindex the sets such that if $W^u(\Lambda_i) \cap W^s(\Lambda_j) \neq \emptyset$, then $i \le j$.

Now, we use the method of Robinson [28] to construct compatible families of unstable disks.

LEMMA 4. *There are neighborhoods U_i of Λ_i and families $\{Z^{u\varphi}_{ix} : x \in O(U_i)\}$, where each $Z^{u\varphi}_{ix}$ is a C^1 disk in η of dimension equal dimension $\eta^u|\Lambda_i$ such that*

1. *$F(Z^{u\varphi}_i) \supset Z^{u\varphi}_i$.*
2. *If $x \in W^u(\Lambda_i, \varphi^T)$, then $\exp Z^{u\varphi}_{ix}$ is a local unstable manifold of x.*
3. *$\{Z^{u\varphi}_{ix} : x \in U_i\}$ can be written as the image of $\sigma^{u\varphi} : \eta^u(r_2) \to \eta$ and the slope of $\sigma^{u\varphi}$ is uniformly bounded, where $r_2 > 0$ is a constant.*
4. *If $i \le j$ and $x \in O(U_i) \cap O(U_j)$, then $Z^{u\varphi}_i \supset Z^{u\varphi}_j$.*
5. *Given the d_φ metrics on M and TM, $\sigma^{u\varphi}$ is Lipschitz with a uniform Lipschitz constant as a map from $\eta^u(r_2)$ to η and the Lipschitz jet on the whole bundle varies uniformly continuously along fibers as explained in* [28].

Proof. We construct the families of unstable disks for F as in section 5 of [28] (see also [27]). By induction, assume that they are constructed for $i = 1, \ldots, k-1$ and then construct them for a neighborhood of Λ_k. Let $U \subset V_1$ be a small neighborhood of Λ_k. Let $B^s_k \subset W^s(\Lambda_k)$ be a fundamental domain of φ^T. Take the sets P_{iq} so that $P^{iq} = \bigcup_{p=i}^{k-1} P_{pq}$ is a closed neighborhood of $B^s_k \cap \bigcup_{p=i}^{k-1} W^u(\Lambda_p)$. The section $\sigma_0 : \eta^u(r)|P^{11} \to \eta$ is constructed to be compatible with $\{Z^{u\varphi}_i(x) : x \in O(U_i)\}$ for $1 \le i \le k-1$ and to be d_φ-Lipschitz. Let $U_k = (W^u(\Lambda_k) \cap U) \cup O^+(P^{11}, U)$, where $O^+(P^{11}, U) = \{\varphi^s(x) : x \in P^{11}$ and $\varphi^s(x) \in U$ for $0 \le s \le t\}$. We extend σ_0 to U_k using F as follows. Let $\Sigma(1, r, \sigma_0) = \{$section $\sigma : \eta^u(r)|U_k \to \eta(r)$ such that $\sigma = \sigma_0$ on the domain of σ_0 and *slope*$(\sigma_x) \le 1$ for $x \in U_k\}$. Define $F_\#(\sigma)$ to be the graph

transform of σ by F extended back over P^{11} using σ_0,

$$\begin{cases} F(image(\sigma_{\varphi^{-T}(x)})) \supset image(F_{\#}(\sigma)_x) & \text{if} \quad \varphi^{-T}(x), \qquad x \in U_k, \\ F_{\#}(\sigma_x) = \sigma_{0x} & \text{if} \quad x \in P^{11}. \end{cases}$$

Using Theorems 3.1 and 3.2 of [28], we get that $F_{\#}$ is a contraction on the space of sections $\Sigma(1, r_2, \sigma_0)$ for some sufficiently small constant $r_2 > 0$ and so has a unique fixed point $\sigma^{u\varphi}$ which is d_φ-Lipschitz. Let $Z_{ix}^{u\varphi} = image(\sigma_x^{u\varphi})$ for $x \in U_i$. In this way, we complete the induction step to construct compatible families of unstable disks. □

Step 4 (construction of conjugacy and reparametrization). In order to prove structural stability, we do the same construction for $(N^{\frac{T}{n}})^n$. Because $(N^{\frac{T}{n}})^n \to \varphi^T$ in the C^1 topology as $n \to \infty$, we can take n sufficiently large so that $(N^{\frac{T}{n}})^n \in \mathcal{U}$, the neighborhood of φ^T in $\mathrm{Diff}(M)$ where Θ is defined. Let

$$G_n(v_x) \equiv \Theta(\tau(v_x, (N^{\frac{T}{n}})^n), v_x, (N^{\frac{T}{n}})^n) = \exp^{-1}_{\varphi^\tau(x)} \circ (N^{\frac{T}{n}})^n \circ \exp v_x.$$

Then G_n is a C^1 bundle map on $\eta(r_1)$ and C^1 close to F. By the permanence results in Theorem 3.1 of [28] and Theorem 6.1 of [16], we can construct compatible families of unstable disks for G_n over the $O(U_i)$, $\{Z_{ix}^{uN} : x \in O(U_i)\}$. Since G_n^{-1} is a fiber contraction on the bundle $\{Z_{ix}^{uN} : x \in O(U_i)\}$, by induction on i ($i = m, m-1, \ldots, 1$), we can construct a section $v_n : M \to \eta(r_2)$ such that $G_n \circ v_n(x) = v_n \circ \varphi^\tau(x)$. Also v_n is continuous and uniformly d_φ-Lipschitz. Moreover, as n tends to infinity, v_n converges to the zero section in the C^0 and d_φ-Lipschitz senses.

We now define the conjugacy and the reparametrization. Let $H_n(x) = \exp v_n(x)$ and $\tau_n(x) = \tau(v_n(x), (N^{\frac{T}{n}})^n)$ for all $x \in M$. Then both $H_n : M \to M$ and $\tau_n : M \to \mathbb{R}$ are continuous, and H_n converges to the identity in the C^0 sense as $n \to \infty$. Since v_n is invariant by $G_{n\#}$, we get that $v_n \circ \varphi^{\tau_n(x)}(x) = \exp^{-1}_{\varphi^{\tau_n(x)}(x)} \circ (N^{\frac{T}{n}})^n \circ \exp v_n(x)$ and so $H_n \circ \varphi^{\tau_n(x)}(x) = (N^{\frac{T}{n}})^n \circ H_n(x)$. Because H_n is homotopic to the identity, H_n is onto on M. Last, we prove that H_n is one to one in the following lemma and hence complete the proof of Theorem 1.

LEMMA 5. *For all n sufficiently large, H_n is one to one.*

Proof. Let $\tau_n(1, x) = \tau_n(x)$ and $\tau_n(i+1, x) = \tau_n(1, \varphi^{\tau_n(i,x)}(x))$ for $i \geq 1$. If $H_n(x) = H_n(y)$, then $d(x, y) \leq d(x, H_n(x)) + d(H_n(y), y) \leq 2r_2$ and $H_n(\varphi^{\tau_n(i,x)}(x)) = ((N^{\frac{T}{n}})^n)^i(H_n(x)) = ((N^{\frac{T}{n}})^n)^i(H_n(y)) = H_n(\varphi^{\tau_n(i,y)}(y))$ for all $i \in \mathbb{N}$. As Robinson pointed out in the proof of Theorem 2.3 of [27], there exist i_0 and $B > 0$ such that $d_\varphi(p, q) \leq 2Bd(p, q)$, where $p = \varphi^{\tau_n(i_0,x)}(x)$ and $q = \varphi^{\tau_n(i_0,y)}(y)$. By Lemma 2.3 of [25], there is a constant $\alpha > 0$ such that

$$\alpha d(p, q) - |v_n(p) - v_n(q)| \leq d(\exp(v_n(p)), \exp(v_n(q))).$$

Thus

$$\begin{aligned} 0 &= d(H_n(p), H_n(q)) \\ &\geq \alpha d(p, q) - |v_n(p) - v_n(q)| \\ &\geq \alpha d(p, q) - \Lambda(v_n) d_\varphi(p, q) \\ &\geq \alpha d(p, q) - \Lambda(v_n) 2Bd(p, q) \\ &\geq (\alpha - \Lambda(v_n) 2B) d(p, q). \end{aligned}$$

Because $\Lambda(v_n) \to 0$ as $n \to \infty$, we can take n large enough so that $\alpha - \Lambda(v_n)2B > 0$. Then $d(p,q) = 0$ and $p = q$. Therefore x and y are on the same orbit of φ^t. Exponentiating $\eta(r_2)|_{O(x)}$ forms transversal disks along the orbit $O(x)$. But transversal disks are disjoint for two nearby points on the same orbit. Thus $x = y$. This shows that H_n is one to one. □

Remark. It is not difficult to use the above argument to prove Theorem 3. Note that, in this case, the space of the vector field restricted to the chain recurrent set is the zero space, i.e., $Span(X)|\mathcal{R}_\varphi = \{0\}$, and so reparametrization is not necessary. We can simply define our bundle maps F and G_n by $F(v_x) = \exp^{-1}_{\varphi^T(x)} \circ \varphi^T \circ \exp v_x$ and $G_n(v_x) = \exp^{-1}_{\varphi^T(x)} \circ (N^{\frac{T}{n}})^n \circ \exp v_x$, then go through the same argument.

Acknowledgments. The author thanks Professor Clark Robinson of Northwestern University for helpful conversation and unknown referees for valuable suggestions.

REFERENCES

[1] D. V. Anosov, *Geodesic Flows on Closed Riemann Manifolds with Negative Curvature*, Proceedings of the Steklov Institute of Mathematics 90, 1967; translated from the Russian by S. Feder, AMS, Providence, RI, 1969.
[2] K. E. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed., John Wiley & Sons, New York, 1989.
[3] W.-J. Beyn, *On invariant curves for one-step methods*, Numer. Math., 51 (1987), pp. 103–122.
[4] W.-J. Beyn, *On the numerical approximation of phase portraits near stationary points*, SIAM J. Numer. Anal., 24 (1987), pp. 1095–1113.
[5] M. Braun, *Differential Equations and Their Applications, 3rd ed.*, Springer-Verlag, New York, 1983.
[6] M. Braun and J. Hershenov, *Periodic solutions of finite difference equations*, Quart. Appl. Math., 35 (1977), pp. 139–147.
[7] H. W. Broer, H. M. Osinga, and G. Vegter, *Algorithms for computing normally hyperbolic invariant manifolds*, Z. Angew. Math. Phys., 48 (1997), pp. 480–524.
[8] H. T. Doan, *Invariant curves for numerical methods*, Quart. Appl. Math., 3 (1985), pp. 385–393.
[9] T. Eirola, *Invariant curves of one-step methods*, BIT, 28 (1988), pp. 113–122.
[10] B. M. Garay, *Discretization and some qualitative properties of ordinary differential equations about equilibria*, Acta Math. Univ. Comen., 62 (1993), pp. 249–275.
[11] B. M. Garay, *Discretization and normal hyperbolicity*, Z. Angew. Math. Mech., 74 (1994), pp. T662–T663.
[12] B. M. Garay, *On structural stability of ordinary differential equations with respect to discretization methods*, Numer. Math., 72 (1996), pp. 449–479.
[13] B. M. Garay, *The discretized flow on domains of attraction: A structural stability result*, IMA J. Numer. Anal., 18 (1998), pp. 77-90.
[14] A. Hagen, *Hyperbolic Structures of Time Discretizations and the Dependence on the Time Step*, Ph.D. thesis, University of Minnesota, Minneapolis, MN, 1996.
[15] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations. Nonstiff Problems,* I, Springer Ser. Comput. Math. 8, Springer-Verlag, New York, 1987.
[16] M. Hirsch, C. Pugh, and M. Shub, *Invariant Manifolds*, Lecture Notes in Math. 583, Springer-Verlag, New York, 1977.
[17] M.-C. Li, *Structural stability of flows under numerics*, J. Differential Equations, 141 (1997), pp. 1–12.
[18] M.-C. Li, *Structural stability of Morse–Smale gradient-like flows under discretizations*, SIAM J. Math. Anal., 28 (1997), pp. 381–388.
[19] M.-C. Li, *Structural stability on basins for numerical methods*, in Proc. Amer. Math. Soc., 127 (1999), pp. 289–295.
[20] J. Lorenz, *Numerics of Invariant Manifolds and Attractors*, in Chaotic Numerics, P. E. Kloeden and K. J. Palmer, eds., Contemp. Math. 172, AMS, Providence, RI, 1993, pp. 185–202.
[21] J. Moser, *On a theorem of Anosov*, J. Differential Equations, 5 (1969), pp. 411–440.
[22] J. Palis and S. Smale, *Structural stability theorems*, in Global Analysis, Proc. Sympos. Pure Math. 14, AMS, Providence, RI, 1970, pp. 223–231.

[23] S. Yu. Pilyugin, *Shadowing in structurally stable flows*, J. Differential Equations, 140 (1997), pp. 238–265.
[24] C. Pugh and M. Shub, *C^r stability of periodic solutions and solution schemes*, Appl. Math. Lett., 1 (1988), pp. 281–285.
[25] J. Robbin, *A structural stability theorem*, Ann. of Math. (2), 94 (1971), pp. 447–493.
[26] C. Robinson, *Structural stability of vector fields*, Ann. of Math. (2), 99 (1974), pp. 154–175.
[27] C. Robinson, *Structural stability of C^1 flows*, in Dynamical Systems—Warwick 1974, Lecture Notes in Math. 468, Springer-Verlag, Berlin, 1975, pp. 262–277.
[28] C. Robinson, *Structural stability of C^1 diffeomorphisms*, J. Differential Equations, 22 (1976), pp. 28–73.
[29] C. Robinson, *Structural stability on manifolds with boundary*, J. Differential Equations, 37 (1980), pp. 1–11.
[30] C. Robinson, *Dynamical Systems: Stability, Symbolic Dynamics, and Chaos*, CRC Press, Boca Raton, FL, 1995.
[31] M. Shub, *Some remarks on dynamical systems and numerical analysis*, in Dynamical Systems and Partial Differential Equations, Proc. VII Elam, Equinoccio, Universidad Simon Bolivar, Caracas, 1986, pp. 69–91.
[32] M. Shub and S. Smale, *Computational complexity: On the geometry of polynomials and a theory of cost: Part* I, Ann. Sci. École. Norm. Sup. (4), 18 (1985), pp. 107–142.
[33] S. Smale, *On gradient dynamical systems*, Ann. of Math. (2), 74 (1961), pp. 199–206.
[34] S. Smale, *Differentiable dynamical systems*, Bull. Amer. Math. Soc., 73 (1967), pp. 747–817.
[35] S. Smale, *The fundamental theorem of algebra and complexity theory*, Bull. Amer. Math. Soc. (N. S.), 4 (1981), pp. 1–36.
[36] A. M. Stuart, *Numerical analysis of dynamical systems*, in Acta Numerica, A. Iserles, ed., Cambridge University Press, Cambridge, UK, 1994, pp. 467–572.
[37] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 756–786

THE CRYSTALLINE VERSION OF THE MODIFIED STEFAN PROBLEM IN THE PLANE AND ITS PROPERTIES*

PIOTR RYBKA†

Abstract. We study the modified Stefan problem in the plane for polygonal interfacial curves. Uniqueness of local in time solutions is shown while existence of local in time solutions has been proved in an earlier work of the author [P. Rybka, *Advances in Differential Equations*, 3 (1998), pp. 687–713]. Geometric properties of the flow are studied if the Wulff shape is a regular N-sided polygon and the initial interface has sufficiently small perimeter. Namely, if the isoperimetric quotient of the initial interface does not differ much from the isoperimetric quotient of the Wulff shape, then the interface shrinks to a point in finite time and the isoperimetric quotient decreases.

Key words. free boundary, Stefan problem, Gibbs–Thomson law, crystalline anisotropy

AMS subject classifications. 35R35, 73B40, 80A22, 52B60

PII. S0036141097325435

1. Introduction. We study a version of the modified Stefan problem in the plane. The special feature of our approach is that we assume that the interfacial curve is a polygon. We stress that admitting nonsmooth interfaces is natural from the viewpoint of modelling crystal evolution. We shall pursue this direction.

Describing the process of melting or growing a crystal requires setting the problem in the framework of two-phase thermodynamics. This is done in the book of Gurtin [Gu]. The author of the book pays special attention to the evolution of nonsmooth interfaces (see Section 12 in [Gu]). Developing this theory Gurtin and Matias proposed in their paper [GM] the particular problem we study here. The setting is the following: a crystal $\Omega_1(t)$ is in a bounded container Ω filled with melt $\Omega_2(t)$, i.e., $\Omega_2(t) = \Omega \backslash \overline{\Omega_1}(t)$ (the notation shall be explained in detail in the next section). The heat transport is described by the equation

$$e_i u_t = -\operatorname{div} \mathbf{q} \quad \text{in} \bigcup_{0<t<T} \Omega_i(t), \quad i = 1, 2, \tag{1.1}$$

which is complemented by the Fourier law

$$\mathbf{q} = -k_i \nabla u, \quad i = 1, 2. \tag{1.2}$$

The temperature u is continuous across the interface $s = \partial\Omega_1 \cap \partial\Omega_2$ being a polygon with facets s_i, $s = \bigcup_{i=1}^{N} s_i$. We assume the number of facets is constant. We are fully aware that this assumption can be the subject of discussion. (We will make some comments on this issue in section 5.) We claim, however, that in the case of small perimeter of $s(t)$ the stabilizing forces of surface tension prevail over destabilizing bulk forces and as a consequence $s(t)$ shrinks to a point. We will prove this in the last section.

*Received by the editors August 4, 1997; accepted for publication (in revised form) July 24, 1998; published electronically May 7, 1999. This work was carried out with the support of the Alexander von Humboldt Foundation at TU München. The author was also partially supported by KBN grant 2 P03A 034 08.

http://www.siam.org/journals/sima/30-4/32543.html

†Institute of Applied Mathematics, Warsaw University, ul. Banacha 2, 02-097 Warszawa, Poland (rybka@mimuw.edu.pl).

Continuing the description of our problem we denote by V_i the velocity of s_i in the direction of the outer normal ν_i. This velocity satisfies the equation

$$[\![\mathbf{q}]\!]\nu_j = V_j, \quad j = 1, \dots, N. \tag{1.3}$$

Finally, we need a condition that the temperature u must satisfy on the interface. The approximation to the balance of capillary forces yields

$$\int_{s_j(t)} u = \Gamma_j - \beta_j L_j(t) V_j(t), \quad j = 1, \dots, N. \tag{1.4}$$

We remark here that the above problem was formulated by Herring in the metallurgical literature in the 1950s; see [Hr]. Later, it was independently rediscovered by Ben Amar–Pomeau [BP] and Gurtin–Matias [GMl].

One might expect here a version of Gibbs–Thomson law (with or without kinetic undercooling). This law states that the temperature on the interface is proportional to the curvature of the interface. In our consideration u is a normalized temperature, so u is zero at melting temperature on a flat interface. Indeed, (1.4) is a version of the Gibbs–Thomson law, where because of lack of smoothness the definition of curvature is adjusted so that it is well defined for polygons. As a matter of fact Γ_i/L_i is the "weighted crystalline curvature" of the edge s_i, $i = 1, \dots, N$ (Γ_i is an appropriate constant).

We note here that the modified Stefan problem for smooth interfaces (without kinetic undercooling) has already been studied. The first paper is by Luckhaus [L], who considered weak solutions and C^1-interfaces. Almgren and Wang [AW] studied this problem in a more general setting allowing for anisotropic surface energy densities.

The presence of nonzero kinetic undercooling makes the analysis somewhat simpler. The modified Stefan problem with kinetic undercooling was studied in particular by Chen and Reitich (see [CR]). They showed local in time existence and uniqueness of smooth temperature u (away from the interface) as well as smooth interface. Independently, Radkevich [Ra] studied the same problem. The advantage of [Ra] is that the author allows a slightly more general form of the heat equation. These results were extended by Soner [So]. He showed global in time existence of weak solutions (his interface is just $(n-1)$-rectifiable). He constructs them as limits of solutions to a phase field model.

For the sake of simplicity we work assuming that the bulk specific heats e_i are equal $e_1 = e_2 = \epsilon > 0$, and similarly we set the coefficients of conductivity k_i, $i = 1, 2$, to be 1. Existence of weak solutions to (1.1)–(1.4) with these simplifications and augmented with initial and boundary conditions has already been established by the author in [Ry2]. However, the method of proof does not yield uniqueness. On the other hand uniqueness holds for the smooth counterpart of our problem. We show that it is also true here. We do this in a way similar to that employed in [Ry1] to show uniqueness for the quasi-steady approximation of (1.1)–(1.4), i.e., for $e_1 = e_2 = 0$. Namely, we derive more or less explicit representation of solutions enabling us to reduce the problem to a uniqueness question for an integral equation. We also point out that unlike the case $e_1 = e_2 = 0$ we crucially depend on all β's being positive, not just nonnegative.

Thus, we touched the question of how (1.1)–(1.4) is related to its quasi-steady approximation. In particular, the problem of convergence of solutions u^ϵ of (1.1)–(1.4) as ϵ goes to zero is open. We leave this for further investigations.

For convenience of the reader we recall in the next section the weak formulation of the problem (1.1)–(1.4) augmented with initial and boundary conditions. In section 3 we derive the representation of solutions which permits us to reduce the problem to a simpler question regarding integral equations. We subsequently prove uniqueness of solutions constructed in [Ry2]. We stress that in any case we have no ground to claim that solutions are global in time. To the contrary, we anticipate that geometry may change during the evolution. In particular for small s_0 we expect finite extinction time.

Apart from showing uniqueness we will present some properties of the flow. The aim of section 4 is to prepare some geometric background needed in the last section. We show in particular that if the isoperimetric ratio L^2/A is only slightly bigger than that of the Wulff shape then the quotient

$$\frac{\max_{i=1,\dots,N} L_i}{\min_{i=1,\dots,N} L_i}$$

remains bounded. Thus, we can improve the qualitative results of [Ry1] for the quasi-steady flow; see section 5.

In order to derive properties of the flow (1.1)–(1.4) we compare it to the system of motion by crystalline curvature

$$\Gamma_i = \beta_i L_i V_i, \tag{1.5}$$

which may be seen as the "zero-temperature-limit" of (1.1)–(1.4). Interestingly, both flows behave similarly if the initial interface has small perimeter.

The system (1.5) was first studied by J. Taylor [T]. Estimates for solutions are particularly easy if the initial polygon is the Wulff shape. But our indispensable tool is a comparison principle of [GG] for solutions of (1.5). We will also use the results of Stancu [St], who derived properties of (1.5) analogous to that for the smooth motion by curvature.

In section 5 we investigate properties of solutions if an initial interface has small perimeter and isoperimetric ratio not much bigger than for the Wulff shape. If the initial temperature distribution is negative and small in an appropriate sense, then the interface shrinks to a point in finite time. We have proved a weaker version of this result in [Ry2, Theorem 4.1], where we assumed that the initial interface is a scaled Wulff shape. Finally, we prove an a priori bound for temperature in the $L^\infty(\Omega)$ norm. This means that no matter how small $s(t)$ is, nor how fast its facets move, the temperature may not drop too much.

2. Weak formulation. Before stating the problem (1.1)–(1.4) in the weak form let us explain the setting and our basic assumptions. We shall consider only *admissible* polygonal interfaces, where the edges s_i and vertices v_i are numbered counterclockwise. Admissibility means here that the outer normals ν_i to the facets s_i belong to the set $\mathcal{S}$ of normals of a given Wulff shape W (cf. Sections 7 and 12 in [Gu]). Moreover, we require that normals to successive facets in s must be neighboring normals to W. For the sake of present analysis we may think of W as being a given, convex polygon with M edges numbered counterclockwise. Let us note that $N \geq M$ and the equality holds if s is convex.

The length of facet s_i determined by its vertices v_i, v_{i+1} is denoted by L_i, $L_i = |v_i - v_{i+1}|$. The perimeter L of s is equal to $\sum_{i=1}^N L_i$. The velocity of V_i of the edge

s_i in the direction of the outer normal ν_i is defined by

$$V_i(t) = \frac{d}{dt} z_i(t),$$

where

$$z_i(t) = \begin{cases} \operatorname{dist}(l_i(t), l_i(0)) & \text{if } (v_i(t) - v_i(0)) \cdot \nu_i > 0, \\ -\operatorname{dist}(l_i(t), l_i(0)) & \text{if } (v_i(t) - v_i(0)) \cdot \nu_i < 0, \end{cases} \tag{2.1}$$

and $l_i(t)$ is the line containing $s_i(t)$. The definition of V_i in (1.3) involves the jump $[\![\cdot]\!]$ across $s(t)$. This quantity is given by

$$[\![\phi]\!](x_0) = \lim_{\Omega_2(t) \ni x \to x_0} \phi(x) - \lim_{\Omega_1(t) \ni x \to x_0} \phi(x), \quad x_0 \in s(t) = \partial\Omega_1(t) \cap \partial\Omega_2(t).$$

We assume that the sets Ω, $\Omega_1(t)$, $\Omega_2(t)$ are regions in $\mathbb{R}^2$, where $\Omega_1(t) \subset\subset \Omega$ and $\Omega = \Omega_1(t) \cup s(t) \cup \Omega_2(t)$. At last we assume that the boundary $\partial\Omega$ of Ω is smooth and Ω is bounded.

The kinetic coefficients $\beta_j > 0$ are constants, so are Γ_j, $j = 1, \dots, N$, and they are defined depending on s as follows (see Section 12.5 in [Gu]):

$$\Gamma_j = \begin{cases} -\ell_j & \text{if } s \text{ is locally convex near both vertices } v_i,\ v_{i+1}, \\ \ell_j & \text{if } s \text{ is locally concave near both vertices } v_i,\ v_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$

where ℓ_j is the length of the edge of the Wulff shape with normal ν_j.

Interestingly, Γ_j / L_j is weighted crystalline curvature of s_j. The relevant definition, which does not need any differential structure of s is given in [T]. We will recall it in a simple setting. But let us first remark that Γ_j is closely related to the underlying interface energy density f (which is basically defined on the unit circle). This is due to the fact that f enters the definition of the Wulff shape W (see Section 7 of [Gu] and especially Section 7.5). It follows from this definition that if d_i is the distance from the origin to the ith edge of W, then

$$d_i = f(\nu_i),$$

where ν_i is the outer normal to the ith edge of W. These relations have no implication for existence theory as well as for the uniqueness result presented here in section 3. On the other hand they are quite important for our geometrical considerations. For the sake of simplicity of exposition we carry out our analysis in sections 4 and 5 assuming that W is a regular polygon, i.e., $d_i = d$, $i = 1, \dots, M$; hence $f(\nu)$ has the same value for all $\nu \in \mathcal{S}$. Of course f must not be constant on $\{|\nu| = 1\}$.

Let us now recall the definition of crystalline curvature (see [T, p. 423]). If z_i is as defined above and $\mathbf{z} = (z_1, \dots, z_N)$, i.e., $s(\mathbf{z})$ is a polygon resulting from s by moving entire facet s_i by z_i in the direction of the normal ν_i, $A(\mathbf{z})$ is the area surrounded by $s(\mathbf{z})$, and $L(\mathbf{z})$ is the perimeter of $s(\mathbf{z})$, then the *crystalline curvature* $\mathcal{K}_i$ *of* s_i is

$$\mathcal{K}_i = -\lim_{\Delta z_i \to 0} \frac{L(\mathbf{z} + \mathbf{e}_i \Delta z_i) - L(\mathbf{z})}{A(\mathbf{z} + \mathbf{e}_i \Delta z_i) - A(\mathbf{z})},$$

where $\mathbf{e}_i$, $i = 1, \dots, N$, are the standard unit vectors of the coordinate axis in $\mathbb{R}^N$. This limit may be evaluated with the aid of the lemma below whose proof we leave to the reader (cf. also [Ry1]).

LEMMA 2.1. *Let us suppose we are given a polygon s with its edges s_i numbered counterclockwise, $i = 1, \dots, N$. If L_i is the length of edge s_i and θ_i is the (oriented) angle between normals ν_{i-1} and ν_i to s_{i-1} and s_i, respectively, then*

$$\begin{aligned} \Delta L_i &:= L_i(\mathbf{z} + \Delta\mathbf{z}) - L_i(\mathbf{z}) \\ &= -\Delta z_i(\operatorname{ctan}\theta_i + \operatorname{ctan}\theta_{i+1}) + \frac{\Delta z_{i-1}}{\sin\theta_i} + \frac{\Delta z_{i+1}}{\sin\theta_{i+1}}, \end{aligned} \tag{2.2}$$

where $\Delta\mathbf{z} = (\Delta z_1, \dots, \Delta z_N)$. If we further assume that s is a convex polygon, the origin belongs to the region bounded by s and d_i is the distance from the origin to s_i, then

$$L_i = -d_i(\operatorname{ctan}\theta_i + \operatorname{ctan}\theta_{i+1}) + \frac{d_{i-1}}{\sin\theta_i} + \frac{d_{i+1}}{\sin\theta_{i+1}}.$$

Here, by convention $s_{N+1} = s_1$, etc. □

We note that for convex polygons Lemma 2.1 implies that

$$\mathcal{K}_i = \frac{\kappa_i}{L_i} < 0, \quad i = 1, \dots, N = M,$$

where

$$\kappa_i = -(\operatorname{ctan}\theta_i + \operatorname{ctan}\theta_{i+1}) + \frac{1}{\sin\theta_i} + \frac{1}{\sin\theta_{i+1}} < 0. \tag{2.3}$$

In the case of convex, admissible polygon s and the Wulff shape W being a regular N-gon formula (2.3) reduces to

$$\kappa_i = \kappa = 2\operatorname{ctan}\theta - \frac{2}{\sin\theta} = -2\tan\frac{\pi}{N}, \tag{2.4}$$

because $\theta_i = \theta = \frac{2\pi}{N}$, $i = 1, \dots, N$. In the special case of s being the Wulff shape we have

$$\ell_i = |\kappa| d, \quad i = 1, \dots, N. \tag{2.5}$$

In order to obtain a closed system we augment (1.1)–(1.4) with initial and boundary data. We consider here only homogeneous Dirichlet boundary data:

$$u(0, x) = u_0(x), \quad s(0) = s_0, \quad u|_{\partial\Omega} = 0 \qquad \text{for } t \geq 0. \tag{2.6}$$

This choice gives us some technical advantages. We shall not consider the Neumann condition, which is physically relevant, because our tools do not apply directly to it.

The process of multiplication (1.1) by a test function, then integration by parts using (1.3) leads to the following definition: a pair $(\mathbf{z}, u)$, where $\mathbf{z}$ is as in (2.1), is called a *weak solution* to (1.1)–(1.4) *and* (2.6) *on* $[0, T)$; if $\mathbf{z} \in C^1([0, T); \mathbb{R}^N)$, $\mathbf{z}(0) = 0$, $u \in C^\alpha([0, T), H_0^1(\Omega))$ with $u(0) = u_0$, $u_t \in L^\infty_{loc}([0, T), H^{-1}(\Omega))$, where $H^{-1}(\Omega)$ is the dual of $H_0^1(\Omega)$, and the identities

(2.7a)

$$\epsilon\langle u_t, h\rangle = -\int_\Omega \nabla u(t, x) \cdot \nabla h(x) dx + \sum_{j=1}^N \int_{s_j(t)} V_j(t) h(x) dl \qquad \text{for all} \quad h \in H_0^1(\Omega),$$

$$\int_{s_j(t)} u\,dl = \Gamma_j - \beta_j L_j(t) V_j(t), \quad j = 1, \dots, N, \tag{2.7b}$$

hold, where $\langle \cdot, \cdot \rangle$ is the pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

Existence of a weak solution $(\mathbf{z}, u)$ on a maximal interval of existence $[0, T_max)$ has been shown in [Ry2]. Here, in section 3 we show uniqueness. We stress that a global existence result cannot be expected, especially if we fix the number of edges, because topological catastrophes like self-intersection, collapsing of a facet to a point, bumping into the boundary are imminent. In the last section we study some qualitative properties of solutions in the case where the Wulff shape is a regular polygon.

Notation. Throughout the paper vector quantities are set in bold, e.g., $\mathbf{z} = (z_1, \dots, z_N)$, the inner product in $\mathbb{R}^k$ is denoted by dot: $\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^k a_i b_i$, $|\mathbf{a}|$ is the Euclidean norm $|\mathbf{a}|^2 = \mathbf{a} \cdot \mathbf{a}$, and finally (f, g) is the inner product in $H_0^1(\Omega)$, i.e., $(f, g) = \int_\Omega \nabla f(x) \cdot \nabla g(x)\,dx$ and $\|f\|^2 = (f, f)$. Occasionally, in order to avoid confusion we will write $(f, g)_{H_0^1(\Omega)}$ for (f, g).

3. Uniqueness of solutions. In this section we show that solutions to (2.7), which we constructed in [Ry2], are unique. We apply a method similar to that used in [Ry1] for the quasi-steady approximation. Namely, we derive a representation of solutions in terms of a Green function (here for the heat operator). Historically, the construction of an appropriate Green function led to existence theorems; see [LSU]. However, we use here results of [LSU] to prove properties of the Green function. We recall that if u is a sufficiently smooth solution of

$$\begin{aligned} &u_t = \Delta u, \qquad u|_{\partial\Omega} = 0, \\ &u(0, x) = u_0(x), \end{aligned}$$

then u may be represented as

$$u(t, x) = \int_\Omega G(x, y, t) u_0(y)\,dy,$$

where G satisfies $(\partial_t - \Delta_x) G(x, y, t) = \delta_y$, i.e., G is the Green function. Some expressions become more handy if we use the above formula for solutions to the heat equations. In particular, we will need the representation in case u_0 is a measure.

We need to introduce some background for transforming (2.7a) into a more appropriate form. We first rewrite (2.7a) in a unified form. Let us recall that if $g \in H^{-1}(\Omega) (= (H_0^1(\Omega))')$, then there exists $g_i \in L^2(\Omega)$, $i = 1, 2$ such that $g = \sum_{i=1}^2 \frac{\partial g_i}{\partial x_i}$, i.e., the pairing $\langle g, h \rangle$ is given by

$$\langle g, h \rangle = -\int_\Omega \sum_{i=1}^2 g_i \frac{\partial h}{\partial x_i}. \tag{3.1}$$

Let us also recall that the mapping

$$H_0^1(\Omega) \ni f \mapsto -\Delta f \in H^{-1}(\Omega)$$

is an isomorphism of Hilbert spaces. If we now set for any $g \in H^{-1}(\Omega)$

$$F = -\Delta^{-1} g,$$

then we can rewrite (3.1) as

$$\langle g, h\rangle = \langle -\Delta F, h\rangle = \langle -\mathrm{div}\nabla F, h\rangle = \int_\Omega \nabla F(x)\cdot\nabla h(x)\,dx = (F, h).$$

Let us now look at the right-hand side of (2.7a). We note that the mapping

$$H_0^1(\Omega) \ni h \mapsto \int_{s_i} h\,dl =: \delta_{s_i}(h) \in \mathbb{R}$$

is a continuous functional over $H_0^1(\Omega)$. Thus we can now define elements f_i, $i = 1, \dots, N$, as follows:

$$f_i = -\Delta^{-1}\delta_{s_i}. \tag{3.2}$$

Thus, by (3.2) we obtain

$$\int_{s_i} h\,dl = -\langle \Delta f_i, h\rangle = (f_i, h).$$

In the following, since s is defined by $\mathbf{z}$ we will write $s(\mathbf{z})$ as well as $f(\mathbf{z})$ in order to stress this dependence.

Thus, after taking into account the above remarks and after setting

$$U = -\Delta^{-1}u,$$

we obtain that $U \in H^3(\Omega)\cap H_0^1(\Omega)$, $U_t \in H_0^1(\Omega)$. We can rewrite (2.7a) as

$$\epsilon(U_t, h) = (\Delta U, h) + \sum_{j=1}^N V_j\,(f_j(\mathbf{z}), h) \quad \text{for all} \quad h \in H_0^1(\Omega).$$

Therefore, (2.7a) is equivalent to

$$\epsilon U_t = \Delta U + \sum_{j=1}^N V_j f_j(\mathbf{z}), \quad U(0) = U_0 \tag{3.3a}$$

and (2.7b) becomes

$$\frac{dz_i}{dt} = \frac{\Gamma_i + (\Delta U, f_i(\mathbf{z}))}{\beta_i L_i}, \quad z_i(0) = 0, \quad i = 1, \dots, N. \tag{3.3 b}$$

We may now recall the following proposition.

Proposition 3.1 (see [Ry2, Theorem 3.1]). *Let us suppose that $\beta_i > 0$, Γ_i, $i = 1, \dots, N$ are as in section 2, $f_i(\mathbf{z})$ is defined by (3.2), and $\epsilon > 0$. We also assume that s_0 is an admissible polygon, $U_0 = -\Delta^{-1}u_0$ is compatible with the problem (3.3), i.e., $u_0 \in H_0^1(\Omega)$ and*

$$u_0 - \sum_{j=1}^N f_j(0)V_j(0) \in H^2(\Omega)\cap H_0^1(\Omega), \tag{3.4}$$

where $V_i(0)$ is given by (3.3 b),

$$V_i(0) = \frac{\Gamma_i - \int_{s_i(0)} u_0\,dl}{\beta_i L_i(0)}.$$

Then, there exists $T_{\max} > 0$ *and* $\mathbf{z}$, U *solutions of* (3.3) *such that*

$$U \in C^\alpha([0, T_{\max}), H^3(\Omega)), \quad U_t \in C^\alpha([0, T_{\max}), H_0^1(\Omega)),$$
$$z_i \in C^{1,\alpha}([0, T_{\max})), \quad i = 1, \dots, N,$$

for any $\frac{1}{2} > \alpha > 0$.

This proposition immediately yields the following corollary.

COROLLARY 3.2 (see [Ry2, Proposition 3.2]). *If* Γ_i, β_i, ϵ, u_0, *and* s_0 *are as in the previous proposition, then the pair* $(u, \mathbf{z})$, *where* $u = -\Delta U$ *is a weak solution to* (2.7) *on the interval* $[0, T_{\max})$ *and*

$$u \in C^\alpha([0, T_{\max}), H_0^1(\Omega)), \qquad z_i \in C^{1,\alpha}([0, T_{\max})), \quad i = 1, \dots, N. \qquad \square$$

Remark. The condition (3.4) is necessary to obtain claimed smoothness of solutions in time.

It is apparent that the transformed system (3.3) is a parabolic equation coupled to an ODE. We will look at (3.3a) not only from the viewpoint of [LSU] but also from the abstract standpoint of semigroup theory. For this matter our basic reference is [Hn] with the changes made in the Russian translation of the book. Having this in mind we define an operator $\mathcal{A} : D(\mathcal{A}) \subset L^2(\Omega) \to L^2(\Omega)$, by

$$\mathcal{A}u = -\Delta u,$$

where $D(\mathcal{A}) = H^2(\Omega) \cap H_0^1(\Omega)$. Then we have the following result whose proof may be found in [Hn, Section 1.6].

LEMMA 3.3. $\mathcal{A}$ *is self-adjoint; moreover* $\mathcal{A}$ *is positive definite and hence it is a sectorial operator.* $\square$

It follows that the fractional powers X^α of $X = L^2(\Omega)$ are well defined. In particular, it is a well-known fact (see Theorems 1.15.3 and 4.3.3 in [Tr]) that

$$X^{1/2} = H_0^1(\Omega). \tag{3.5}$$

We want to apply to (3.3a) the variation of constant formula [Hn, Theorem 3.2.2]. In order to apply this theorem we need to know that the source term is Hölder continuous in time into $L^2(\Omega)$. We showed in [Ry2] (see Lemmas 3.3, 3.4, and 3.5) that $\mathbf{z} \to f_i(\mathbf{z})$ is Hölder continuous with exponent $\alpha \in (0, \frac{1}{2})$ into X^σ, where $\sigma + \frac{1}{2}\alpha < 3/4$. Proposition 3.1 yields $\mathbf{z} \in C^{1,\alpha}(\Omega, \mathbb{R}^N)$ so that we can write

$$U(t) = e^{\Delta t/\epsilon} U_0 + \frac{1}{\epsilon} \int_0^t e^{\Delta(t-\tau)/\epsilon} \sum_{i=1}^N f_i(\tau) V_i(\tau)\, d\tau. \tag{3.6}$$

We can insert this into (3.3b) to obtain

$$\frac{dz_i}{dt}(t) = \frac{\Gamma_i}{\beta_i L_i} - \frac{1}{\beta_i L_i} \left(e^{\Delta t/\epsilon} u_0 + \frac{1}{\epsilon} \int_0^t \Delta e^{\Delta(t-\tau)/\epsilon} \sum_{j=1}^N V_j(\tau) f_j(\tau)\, d\tau, f_i(t) \right)_{H_0^1(\Omega)}$$
$$= \frac{\Gamma_i}{\beta_i L_i} - \frac{1}{\beta_i L_i} \left(\int_{s_i} e^{\Delta t/\epsilon} u_0 + \frac{1}{\epsilon} \int_{s_i} \int_0^t \Delta e^{\Delta(t-\tau)/\epsilon} \sum_{j=1}^N V_j(\tau) f_j(\tau)\, d\tau \right). \tag{3.7}$$

The integral on the right-hand side is well defined but it is not quite convenient to deal with if we want to prove that it is Lipschitz continuous in $\mathbf{z}$. We cannot simply

interchange the order of operations in the last integrand because $-\Delta f_j = \delta_{s_j}$ is a measure. We find it is more convenient to work with Green functions.

Existence of the Green function and parts (a) and (b) of the lemma below are well established (see [LSU]). Essentially, part (c) is also a well-known fact. It is apparent from the eigenfunction representation of G as in [TS, Chapter VI, Sections 1, 2]. However, we present an independent and more general proof, which does not refer to eigenfunctions of Laplace operator.

LEMMA 3.4. *Let us suppose that $\partial\Omega$ is smooth, $g \in L^p(\Omega)$, $p \in [2, \infty)$. Then, there exists a function $G(x, y, t)$ such that $LG(x, y, t) := (\partial_t - \Delta_x)G(x, y, t) = \delta_y$ and for $t \geq 0$, and*

$$e^{\Delta t} g(x) = \int_\Omega G(x, y, t) g(y)\, dy.$$

Moreover, $G(x, y, t) = K_t(x - y) + H(x, y, t)$, where

(a) $K_t(\xi) = 4^{-1}\pi^{-1}t^{-1}\exp(-|\xi|^2/4t)$ *for* $t > 0$; $K_0(\xi) = 0$;

(b) $H \in C^\infty(\bar{\Omega} \times \Omega \times \mathbb{R}_+)$ *and*

$$\begin{aligned} &\frac{\partial H}{\partial t} - \Delta_x H = 0 \quad \text{in } \Omega; \\ &H(x, y, t) = -K_t(x - y) \quad \text{for } x \in \partial\Omega, \quad y \in \Omega; \\ &H(x, y, 0) = 0 \quad \text{for } x \in \bar{\Omega}, \quad y \in \Omega; \end{aligned}$$

(c) $H(x, y, t) = H(y, x, t)$ *for* $(x, y) \in \Omega \times \Omega$, $t > 0$.

Proof. Existence and smoothness of H will follow from the classical theory of parabolic equation after we check that the compatibility conditions hold. It is the case since

$$\frac{\partial^k}{\partial t^k} K_t(x - y) = 0, \quad k = 0, 1, 2, \ldots,$$

for $t > 0$, and $x \in \partial\Omega$, $y \in \Omega$. We may now invoke Theorem 5.2 of [LSU, Chapter IV] to finish the proof of (b).

Let us now set

$$w(x, t) = \int_\Omega [K_t(x - y) + H(x, y, t)] g(y)\, dy.$$

It is not difficult to check that $Lw = 0$, $w(x, 0) = g(x)$, $w|_{\partial\Omega} = 0$, so by uniqueness of solutions to the heat equations we have

$$e^{\Delta t} g(x) = \int_\Omega G(x, y, t) g(y)\, dy.$$

We still need to prove (c). It is well known that if $\mathcal{A}$ is a self-adjoint positive operator then $e^{-\mathcal{A}t}$ is self-adjoint too (e.g., this follows from the representation for $e^{-\mathcal{A}t}$ given in [Hn, Theorem 1.3.4]). Now, if $T : L^2(\Omega) \to L^2(\Omega)$ is given by

$$(Tf)(x) = \int_\Omega G(x, y) f(y)\, dy,$$

where $G \in L^2(\Omega \times \Omega)$ then it is also well known that

$$(Tf)^*(x) = \int_\Omega \bar{G}(y, x) f(y)\, dy.$$

Thus it follows that

$$\int_\Omega [K_t(x-y)+H(x,y,t)]g(y)\,dy = e^{\Delta t}g(x)$$
$$= ((e^{\Delta t})^* g)(x) = \int_\Omega [\bar{K}_t(y-x) + \bar{H}(y,x,t)]g(y)\,dy.$$

Since $K_t(\xi)$ and $H(x,y,t)$ are real and $K_t(-\xi) = K_t(\xi)$ it follows that

$$\int_\Omega H(x,y,t)g(y)\,dy = \int_\Omega H(y,x,t)g(y)\,dy$$

and finally

$$H(x,y,t) = H(y,x,t),$$

as desired. □

The above lemma much simplifies the form of the RHS of (3.7). We have the following lemma.

LEMMA 3.5. *We assume that f_i is as before, $t > 0$, $s_i \subset\subset \Omega$, then*

$$\Delta e^{\Delta t} f_i(x) = -\int_{s_i} [K_t(x-y) + H(x,y,t)]\,dy.$$

Proof. Let us take the inner product of the LHS with an arbitrary $g \in C(\Omega)$. Because $e^{\Delta t}$ is smoothing for $t > 0$ and self-adjoint we see that

$$\begin{aligned}\int_\Omega \Delta e^{\Delta t} f_i(x)g(x)\,dx &= \int_\Omega \Delta e^{\Delta t/2} e^{\Delta t/2} f_i(x)g(x)\,dx \\ &= \int_\Omega \Delta e^{\Delta t/2} f_i(x) e^{\Delta t/2} g(x)\,dx = \int_\Omega e^{\Delta t/2} f_i(x) \Delta e^{\Delta t/2} g(x)\,dx \\ &= \int_\Omega f_i(x)\Delta e^{\Delta t} g(x)\,dx = -\int_{s_i} e^{\Delta t} g(x)\,dl(x) \\ &= -\int_{s_i}\int_\Omega [K_t(x-y) + H(x,y,t)]g(y)\,dy dl(x),\end{aligned}$$

where we used also the definition of f_i. By part (c) of the previous lemma we have

$$\int_\Omega \Delta e^{\Delta t} f_i(x)g(x)\,dx = -\int_\Omega \int_{s_i} [K_t(x-y) + H(x,y,t)]\,dl(y)g(x)\,dx$$

for all $g \in C(\Omega)$. The lemma follows. □

We immediately apply this result to transform (3.7) further

$$\frac{dz_i}{dt}(t) = \frac{\Gamma_i}{\beta_i L_i} - \frac{1}{\beta_i L_i}\int_{s_i(t)}\int_\Omega G(x,y,t/\epsilon)u_0(y)\,dy dl(x)$$
$$+ \frac{1}{\epsilon\beta_i L_i}\sum_{j=1}^N \int_0^t \int_{s_i(t)}\int_{s_j(\tau)} G(x,y,(t-\tau)/\epsilon)V_j(\tau)\,dl(x)\,dl(y)d\tau.$$

We note that the last integrand is well defined for $\tau \neq t$ and

$$\|H\|_{L^\infty(\Omega_\eta \times \Omega_\eta)} \leq C(\eta),$$

where

$$\Omega_\eta = \{x \in \Omega : \operatorname{dist}(x, \partial\Omega) \geq \eta > 0\}.$$

Let us now state our main theorem of the section.

THEOREM 3.6. *There exists exactly one weak solution of* (2.7).

We recall that we have already shown existence of at least one solution [Ry2, Theorem 3.1]. We will prove here that there is no more than one solution and the above representation suggests the idea of proof. We will show that the difference of velocities $\mathbf{V}$ and $\mathbf{V}'$ corresponding to two solutions satisfies an integral equation. We will be able to conclude our result if we know that the function

$$M_{ij}(\mathbf{z}_1, \mathbf{z}_2, t) := \int_{s_i(\mathbf{z}_1)} \int_{s_j(\mathbf{z}_2)} G(x, y, t)\, dl(y) dl(x)$$

is Lipschitz continuous. It turns out that the Lipschitz constant blows up, but at an integrable (in time) rate. Therefore, we have the following lemma.

LEMMA 3.7. *If $\zeta > 0$, $dist(s(\mathbf{z}_k), \partial\Omega) > \eta > 0$, $k = 1, 2$, then there exists a neighborhood of $(\mathbf{z}_1, \mathbf{z}_2) \in \mathbb{R}^N \times \mathbb{R}^N$ such that for all $(\mathbf{z}_1', \mathbf{z}_2')$ in this neighborhood we have*

$$|M_{ij}(\mathbf{z}_1, \mathbf{z}_2, \zeta) - M_{ij}(\mathbf{z}_1', \mathbf{z}_2', \zeta)| \leq \frac{C}{\sqrt{\zeta}}(|z_1 - z_1'| + |z_2 - z_2'|),$$

where the constant $C = C(\eta, \mathbf{z}_1, \mathbf{z}_2)$ is independent of ζ.

Proof. We will proceed in a few steps. Since $G(x, y, \zeta) = K_\zeta(x - y) + H(x, y, \zeta)$, where $K_\zeta(x-y)$ is singular and away from $\partial\Omega$ the function $H(x, y, \zeta)$ is bounded with its derivative, then we will first look at

$$M_{ij}'(\mathbf{z}_1, \mathbf{z}_2, \zeta) = \int_{s_i(\mathbf{z}_1)} \int_{s_j(\mathbf{z}_2)} K_\zeta(x - y)\, dl(y) dl(x).$$

(a) $s_i(\mathbf{z}_1)$ and $s_j(\mathbf{z}_2)$ are not parallel. If it is so, then the lines containing $s_i(\mathbf{z}_1)$ and $s_j(\mathbf{z}_2)$ intersect at point p. Hence, we can choose a neighborhood of $(\mathbf{z}_1, \mathbf{z}_2)$ in $\mathbb{R}^N \times \mathbb{R}^N$ such that the lines containing $s_i(\mathbf{z}_1')$ and $s_j(\mathbf{z}_2')$ intersect at point p' which is close to p. We will describe these segments precisely in order to facilitate our analysis. According to the notation convention of section 2 v_i and v_{i+1} are vertices of s_i. We define $\mathbf{w}_i = (v_{i+1} - v_i)/|v_{i+1} - v_i|$ and analogously $\mathbf{w}_j = (v_{j+1} - v_j)/|v_{j+1} - v_j|$, furthermore we set

$$a_i = \begin{cases} |v_i - p| & \text{if } \mathbf{w}_i \cdot (v_i - p) \geq 0, \\ -|v_i - p| & \text{if } \mathbf{w}_i \cdot (v_i - p) < 0. \end{cases}$$

We adopt the analogous definitions for a_j, a_i', and a_j'. Hence, we have

$$s_i(\mathbf{z}_1) = \{p + t_i\mathbf{w}_i :\ t_i \in [a_i, a_i + L_i]\}, \quad s_j(\mathbf{z}_2) = \{p + t_j\mathbf{w}_j :\ t_j \in [a_j, a_j + L_j]\},$$
$$s_i(\mathbf{z}_1') = \{p' + t_i\mathbf{w}_i :\ t_i \in [a_i', a_i' + L_i']\}, \quad s_j(\mathbf{z}_2') = \{p' + t_j\mathbf{w}_j :\ t_j \in [a_j', a_j' + L_j']\}.$$

Having this in mind, if $x \in s_i(\mathbf{z}_1')$, $y \in s_j(\mathbf{z}_2')$ (possibly $\mathbf{z}_1' = \mathbf{z}_1$, $\mathbf{z}_2' = \mathbf{z}_2$), then we calculate

$$|x - y|^2 = |t_i\mathbf{w}_i - t_j\mathbf{w}_j|^2 = t_i^2 + t_j^2 - 2\mathbf{w}_i \cdot \mathbf{w}_j t_i t_j \geq (t_i^2 + t_j^2)(1 - \cos\theta) > 0, \tag{3.8}$$

because the angle between $\mathbf{w}_i$ and $\mathbf{w}_j$ is no less than θ. We may now estimate the difference $|M'_{ij}(\mathbf{z}_1,\mathbf{z}_2,\zeta) - M'_{ij}(\mathbf{z}'_1,\mathbf{z}'_2,\zeta)| =: I$ using the definition of $K_\zeta(x-y)$

$$I \le \frac{1}{\pi\zeta}\left|\int_{a_i}^{a_i+L_i}\int_{a_j}^{a_j+L_j} e^{-\frac{|t_i\mathbf{w}_i - t_j\mathbf{w}_j|^2}{4\zeta}}\,dt_i dt_j - \int_{a'_i}^{a'_i+L'_i}\int_{a'_j}^{a'_j+L'_j} e^{-\frac{|t_i\mathbf{w}_i - t_j\mathbf{w}_j|^2}{4\zeta}}\,dt_i dt_j\right|.$$

Let us set

$$Q = [a_i, a_i+L_i]\times[a_j, a_j+L_j], \qquad Q' = [a'_i, a'_i+L'_i]\times[a'_j, a'_j+L'_j].$$

It is now clear that there is no contribution from the integrals over $Q\cap Q'$ to the difference. Therefore we have to estimate the contribution from the symmetric difference $Q\triangle Q' = \bigcup_{k=1}^4 R_k$; see Figure 1. Lemma 2.1 provides us with the estimates for the lengths of edges of R_k, $k=1,2,3,4$. The list of estimates goes as follows:

$$\begin{aligned}
R_1:&\quad |a_j - a'_j| \text{ and } L_i + C|\mathbf{z}_1 - \mathbf{z}'_1|;\\
R_2:&\quad |a_i - a'_i| \text{ and } L_j + C|\mathbf{z}_2 - \mathbf{z}'_2|;\\
R_3:&\quad |a_i - a'_i| + C|\mathbf{z}_1 - \mathbf{z}'_1| \text{ and } L_j + C|\mathbf{z}_2 - \mathbf{z}'_2|;\\
R_4:&\quad |a_j - a'_j| + C|\mathbf{z}_2 - \mathbf{z}'_2| \text{ and } L_i + C|\mathbf{z}_1 - \mathbf{z}'_1|;
\end{aligned}$$

for some constant C if $(\mathbf{z}'_1,\mathbf{z}'_2)$ is sufficiently close to $(\mathbf{z}_1,\mathbf{z}_2)$. We now estimate $|a_j - a'_j|$ as well as $|a_j - a'_j|$. Restricting if necessary the neighborhood of $(\mathbf{z}_1,\mathbf{z}_2)$ we are working in we may assume that

$$a_i a'_i \ge 0 \qquad \text{and} \qquad a_j a'_j \ge 0.$$

Let us set

$$\mathbf{p} = p' - p, \quad \mathbf{u}_i = v'_i - v_i, \quad \mathbf{u}_j = v'_j - v_j,$$

then we see

$$\begin{aligned}
\mathbf{p} &= \nu_i(z'_{1i} - z_{1i}) + \nu_j(z'_{2j} - z_{2j}),\\
\mathbf{u}_i &= \nu_i(z'_{1i} - z_{1i}) + \nu_{i-1}(z'_{1(i-1)} - z_{1(i-1)}),\\
\mathbf{u}_j &= \nu_j(z'_{2j} - z_{2j}) + \nu_{j-1}(z'_{2(j-1)} - z_{2(j-1)}),
\end{aligned} \tag{3.9}$$

where ν_k is the outer normal to s_k, $k=i,j$ (see section 2).

It is easy to observe that if a, b are real and $ab\ge 0$, then $|a-b| = ||a|-|b||$. This observation and the triangle inequality yield

$$|a_i - a'_i| = ||v_i - p| - |v'_i - p'|| \le |\mathbf{u}_i - \mathbf{p}|.$$

By (3.9) we arrive at

$$|a_i - a'_i| \le c(N)|\mathbf{z}_1 - \mathbf{z}'_1| + |\mathbf{z}_2 - \mathbf{z}'_2|),$$

where $c(N)$ depends only on N.

Combining these estimates with (3.8) we may estimate the integral over R_1

$$\begin{aligned}
\frac{1}{\pi\zeta}\int_{R_1} e^{-|t_i\mathbf{w}_i - t_j\mathbf{w}_j|^2/4\zeta}\,dt_i dt_j &\le \frac{1}{\pi\zeta}\int_{R_1} e^{-\frac{1-\cos\theta}{4\zeta}(t_i^2+t_j^2)}\,dt_i dt_j\\
&\le C\frac{|a_i - a'_i|}{\pi\zeta}\int_{-\infty}^{+\infty} e^{-\frac{(1-\cos\theta)}{4\zeta}t_j^2}\,dt_j = \frac{C|a_i - a'_i|}{\sqrt{\pi\zeta}\sqrt{1-\cos\theta}}\\
&\le \frac{C}{\sqrt{\zeta}}(|\mathbf{z}_1 - \mathbf{z}'_1| + |\mathbf{z}_2 - \mathbf{z}'_2|).
\end{aligned}$$

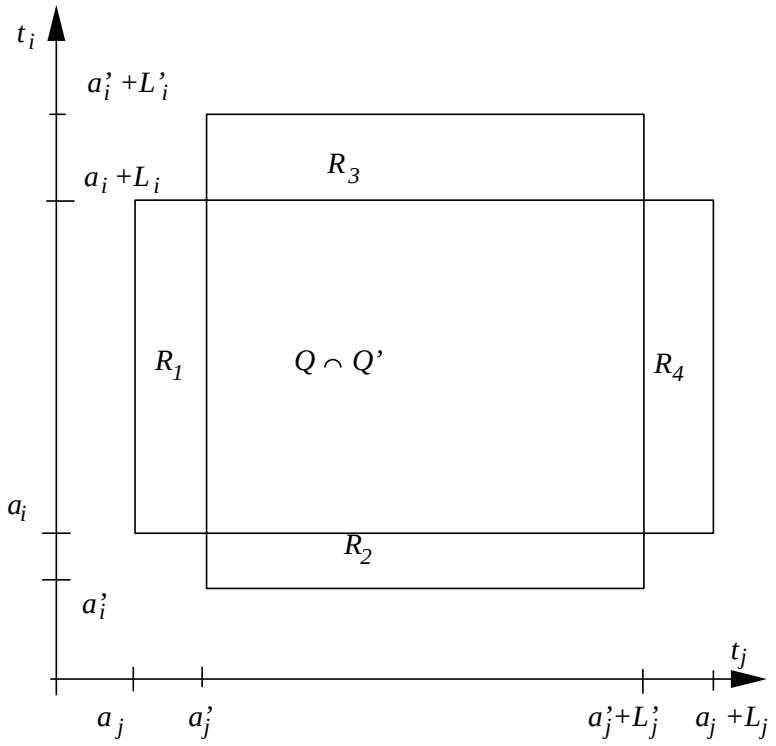


FIG. 1.

Similar calculations lead us to

$$\frac{1}{\pi\zeta}\int_{R_k} e^{-|t_i\mathbf{w}_i - t_j\mathbf{w}_j|^2/4\zeta}\, dt_i dt_j \le \frac{C}{\sqrt{\zeta}}(|\mathbf{z}_1 - \mathbf{z}_1'| + |\mathbf{z}_2 - \mathbf{z}_2'|)$$

for $k = 2, 3, 4$. It follows that

$$|M_{ij}'(\mathbf{z}_1, \mathbf{z}_2, \zeta) - M_{ij}'(\mathbf{z}_1', \mathbf{z}_2', \zeta)| \le C\frac{1}{\sqrt{(1-\cos\theta)\pi\zeta}}(|\mathbf{z}_1 - \mathbf{z}_1'| + |\mathbf{z}_2 - \mathbf{z}_2'|).$$

(b) $s_i(\mathbf{z}_1)$ and $s_j(\mathbf{z}_2)$ are parallel, possibly on the same line. Thus we may write

$$s_i(\mathbf{z}_1) = \{x:\ x = \mathbf{w}t_i + \bar{v}_i, t_i \in [0, L_i]\}, \qquad s_j(\mathbf{z}_2) = \{y:\ y = -\mathbf{w}t_j + \bar{v}_j, t_j \in [0, L_j]\},$$

where $\mathbf{w} = \pm(v_{i+1} - v_i)/|v_{i+1} - v_i|$ and $\bar{v}_k \in \{v_{k+1}, v_k\}$, $k = i, j$. We can choose $\mathbf{w}$, $\bar{v}_i$, and $\bar{v}_j$ in a such a way that

$$\mathbf{w} \cdot (\bar{v}_i - \bar{v}_j) > 0. \tag{3.10}$$

We leave to the reader a proof of this simple geometric fact. We note that after replacing $\bar{v}_k$ with $\bar{v}_k'$, $k = i, j$ inequality (3.10) holds for $(\mathbf{z}_1', \mathbf{z}_2')$ sufficiently close to $(\mathbf{z}_1, \mathbf{z}_2)$. Taking (3.10) into account we obtain for $x \in s_i(\mathbf{z}_1')$, $y \in s_j(\mathbf{z}_2')$ (possibly $\mathbf{z}_1' = \mathbf{z}_1$, $\mathbf{z}_2' = \mathbf{z}_2$)

$$\begin{aligned} |x - y|^2 &= |t_i\mathbf{w} + \bar{v}_i' + t_j\mathbf{w} - \bar{v}_j'|^2 = |(t_i + t_j)\mathbf{w} + (\bar{v}_i' - \bar{v}_j')|^2 \\ &= (t_i + t_j)^2 + |\bar{v}_i' - \bar{v}_j'|^2 + 2(t_i + t_j)\mathbf{w} \cdot (\bar{v}_i' - \bar{v}_j') \ge (t_i^2 + t_j^2). \end{aligned} \tag{3.11}$$

We may now argue as in step (a) using (3.11) in place of (3.8) that

$$|M'_{ij}(\mathbf{z}_1,\mathbf{z}_2,\zeta)-M'_{ij}(\mathbf{z}'_1,\mathbf{z}'_2,\zeta)|\le C\frac{1}{\sqrt{(1-\cos\theta)\pi\zeta}}(|\mathbf{z}_1-\mathbf{z}'_1|+|\mathbf{z}_2-\mathbf{z}'_2|).$$

(c) It remains to show that

$$\left|\int_{s_i(\mathbf{z}_1)}\int_{s_j(\mathbf{z}_2)}H(x,y,\zeta)\,dl(y)dl(x)-\int_{s_i(\mathbf{z}'_1)}\int_{s_j(\mathbf{z}'_2)}H(x,y,\zeta)\,dl(y)dl(x)\right| \le C(|\mathbf{z}_1-\mathbf{z}'_1|+|\mathbf{z}_2-\mathbf{z}'_2|).$$

This part is now easy and we leave it to the reader. □

There is one more term in the integral equation for $\mathbf{V}$ whose Lipschitz continuity has to be investigated.

LEMMA 3.8. *Let us suppose that $\zeta>0$, $u_0\in H^1_0(\Omega)$, and dist$(s_i(z),\partial\Omega)>\eta>0$ then the map*

$$\mathbf{z}\mapsto\int_{s_i(\mathbf{z})}e^{\Delta\zeta}u_0\,dl(y)$$

is locally Lipschitz continuous with the Lipschitz constant independent of ζ.

Proof. We note

$$\begin{aligned}\int_{s_i(\mathbf{z})}e^{\Delta\zeta}u_0\,dl(x)&-\int_{s_i(\mathbf{z}')}e^{\Delta\zeta}u_0\,dl(y)\\ &=\int_{s_i(\mathbf{z})}\int_\Omega K_\zeta(x-y)u_0(y)\,dydl(x)-\int_{s_i(\mathbf{z}')}\int_\Omega K_\zeta(x-y)u_0(y)\,dydl(x)\\ &\quad+\int_{s_i(\mathbf{z})}\int_\Omega H(x,y,\zeta)u_0(y)\,dydl(x)-\int_{s_i(\mathbf{z}')}\int_\Omega H(x,y,\zeta)u_0(y)\,dydl(x)\\ &=I_1+I_2.\end{aligned}$$

We now estimate I_1. We may extend u_0 by zero to the entire plane, so $u_0\in H^1(\mathbb{R}^2)$. We also set $\mathbf{v}=(z'_i-z_i)\nu_i$

$$\begin{aligned}I_1&=\int_{s_i(\mathbf{z})}\int_{\mathbb{R}^2}K_\zeta(x-y)u_0(y)\,dydl(x)-\int_{s_i(\mathbf{z}')}\int_{\mathbb{R}^2}K_\zeta(x-y)u_0(y)\,dydl(x)\\ &=\int_{s_i(\mathbf{z})}\int_{\mathbb{R}^2}K_\zeta(x-y)u_0(y)\,dydl(x)-\int_{s_i(\mathbf{z}')-\mathbf{v}}\int_{\mathbb{R}^2}K_\zeta(x+\mathbf{v}-y)u_0(y)\,dydl(x)\\ &=\int_{s_i(\mathbf{z})\cap(s_i(\mathbf{z}')-\mathbf{v})}\int_{\mathbb{R}^2}K_\zeta(x-y)(u_0(y)-u_0(y-\mathbf{v}))\,dydl(x)\\ &\quad+\int_{s_i(\mathbf{z})\triangle(s_i(\mathbf{z}')-\mathbf{v})}\int_{\mathbb{R}^2}(\chi(x)K_\zeta(x-y)-(1-\chi(x))K_\zeta(x+\mathbf{v}-y))u_0(y)\,dydl(x)\\ &=I_3+I_4,\end{aligned}$$

where

$$\chi(x)=\begin{cases}1 & \text{if } x\in s_i(\mathbf{z})\setminus(s_i(\mathbf{z}')-\mathbf{v});\\ 0 & \text{if } x\in(s_i(\mathbf{z}')-\mathbf{v})\setminus s_i(\mathbf{z}).\end{cases}$$

Schwarz inequality and $\int_{\mathbb{R}^2} K_\zeta^2(x)\,dx = 1$ lead us to

$$\begin{aligned}|I_3| &\le \int_{s_i(\mathbf{z})\cap(s_i(\mathbf{z}')-\mathbf{v})} \|K_\zeta(x-\cdot)\|_{L^2(\Omega)}\|\nabla u_0\|_{L^2(\Omega)}\,dl(x)|\mathbf{v}| \\ &\le |\mathbf{z}-\mathbf{z}'||s_i(\mathbf{z})|\|\nabla u_0\|_{L^2(\Omega)}.\end{aligned}$$

Similar calculations and $|s_i(\mathbf{z})\triangle(s_i(\mathbf{z}')-\mathbf{v})| \le C|\mathbf{z}-\mathbf{z}'|$ imply that

$$|I_4| \le C|\mathbf{z}-\mathbf{z}'|\|u_0\|_{L^2(\Omega)}.$$

Combining estimates for I_3, I_4 we obtain

$$|I_1| \le C|\mathbf{z}-\mathbf{z}'|(\|\nabla u_0\|_{L^2(\Omega)} + \|u_0\|_{L^2(\Omega)}),$$

where the constant C is independent of $\zeta > 0$.

We now estimate I_2. We note

$$\begin{aligned}|I_2| \le &\int_{s_i(\mathbf{z})\cap(s_i(\mathbf{z}')-\mathbf{v})}\int_{\mathbb{R}^2} |H(x,y,\zeta)-H(x+\mathbf{v},y,\zeta)||u_0(y)|\,dydl(x) \\ &+\int_{s_i(\mathbf{z})\triangle(s_i(\mathbf{z}')-\mathbf{v})}\int_{\mathbb{R}^2} |H(x,y,\zeta)-H(x+\mathbf{v},y,\zeta)||u_0(y)|\,dydl(x) = I_5+I_6.\end{aligned}$$

Since $\operatorname{dist}(s_i(\mathbf{z}),\partial\Omega)$, $\operatorname{dist}(s_i(\mathbf{z}'),\partial\Omega) > \eta$, then

$$|H(x,y,\zeta)-H(x+\mathbf{v},y,\zeta)| \le \|D_xH\|_{L^\infty(\Omega_\eta\times\Omega\times[0,\zeta])}|\mathbf{v}|.$$

We claim that $\|D_xH\|_{L^\infty(\Omega_\eta\times\Omega\times[0,\zeta])}$ is independent of ζ. In order to see this we consider the equation defining H (see Lemma 3.4(b)) when $x\in\bar\Omega$ and $y\in\Omega_\eta$. We may differentiate this equation as well as the boundary conditions with respect to y_k, $k=1,2$. We apply the maximum principle to the resulting heat equation. This yields

$$\begin{aligned}\max_{(x,t)\in\Omega\times[0,\zeta]}|D_{y_k}H(x,y,t)| &\le \max_{(x,t)\in\partial\Omega\times[0,\zeta]}\left|\frac{x_k-y_k}{2t}K_t(x-y)\right| \\ &\le \max_{(x,t)\in\partial\Omega\times[0,\zeta]} Ct^{-3/2}e^{-\frac{|x-y|^2}{8t}}\end{aligned}$$

for some C independent of ζ. Since $y\in\Omega_\eta$ we infer that

$$\max_{(x,t)\in\Omega\times[0,\zeta]}|D_{y_k}H(x,y,t)| \le C\max_{t\in[0,\zeta]} t^{-3/2}e^{-\eta^2/8t} \le A < \infty.$$

Hence by Lemma 3.4(c)

$$\|D_xH\|_{L^\infty(\Omega_\eta\times\Omega\times[0,\zeta])} = \|D_yH\|_{L^\infty(\Omega\times\Omega_\eta\times[0,\zeta])} \le A.$$

Let us also note that this argument shows that

$$\|H\|_{L^\infty(\Omega_\eta\times\Omega\times[0,\zeta])} = \|H\|_{L^\infty(\Omega\times\Omega_\eta\times[0,\zeta])} \le A. \tag{3.12}$$

Thus, we easily see that

$$I_5 \le A|v||\Omega|^{1/2}\|u_0\|_{L^2(\Omega)} \le C|\mathbf{z}-\mathbf{z}'|\|u_0\|_{L^2(\Omega)}.$$

The estimate for I_6 is even simpler:

$$\begin{aligned} I_6 &\le |s_i(\mathbf{z}) \triangle (s_i(\mathbf{z}') - \mathbf{v})| \max_x \int_{\mathbb{R}^2} |H(x,y,\zeta) - H(x+\mathbf{v},y,\zeta)||u_0(y)|\, dydl(x) \\ &\le 2C|\mathbf{z}-\mathbf{z}'| \|H\|_{L^\infty(\Omega_\eta \times \Omega \times [0,\zeta])} |\Omega|^{1/2} \|u_0\|_{L^2(\Omega)} \le C'|\mathbf{z}-\mathbf{z}'| \|u_0\|_{L^2(\Omega)}, \end{aligned}$$

where C' is independent from ζ. The lemma follows after we combine the estimates for I_1 and I_2. ∎

We remark that it is important to assume that $\zeta > 0$; without this assumption the lemma is false.

We are now in a position to complete the proof of Theorem 3.6. We have

$$\begin{aligned} V_i = {} & \frac{\Gamma_i}{\beta_i L_i} - \frac{1}{\beta_i L_i} \int_{s_i(t)} \int_\Omega G(x,y,t) u_0(y)\, dydl(x) d\tau \\ & + \frac{1}{\epsilon \beta_i L_i} \sum_{j=1}^N \int_0^t M_{ij}(\mathbf{z}(t), \mathbf{z}(\tau), (t-\tau)/\epsilon) V_j(\tau)\, d\tau \end{aligned} \tag{3.13}$$

for $i = 1, \dots, N$. This identity is valid for any weak solution. Suppose now we have two of them, $(\mathbf{z}', u')$ and $(\mathbf{z}, u)$. We take the difference $V_i - V_i'$. By (3.13), for $t > 0$ we have

$$\begin{aligned} V_i - V_i' = {} & \left(\frac{1}{\beta_i L_i} - \frac{1}{\beta_i L_i'} \right) \left(\Gamma_i + \int_{s_i(t)} \int_\Omega G(x,y,t) u_0(y)\, dydl(x) \right) \\ & + \left(\frac{1}{\epsilon \beta_i L_i} - \frac{1}{\epsilon \beta_i L_i'} \right) \sum_{j=1}^N \int_0^t M_{ij} \left(\mathbf{z}(t), \mathbf{z}(\tau), \frac{t-\tau}{\epsilon} \right) V_j(\tau)\, d\tau \\ & + \frac{1}{\beta_i L_i'} \left(\int_{s_i'(t)} \int_\Omega G(x,y,t) u_0(y)\, dydl(x) - \int_{s_i(t)} \int_\Omega G(x,y,t) u_0(y)\, dydl(x) \right) \\ & + \frac{1}{\epsilon \beta_i L_i'} \sum_{j=1}^N \int_0^t M_{ij} \left(\mathbf{z}(t), \mathbf{z}(\tau), \frac{t-\tau}{\epsilon} \right) (V_j(\tau) - V_j'(\tau))\, dt \\ & + \frac{1}{\epsilon \beta_i L_i'} \sum_{j=1}^N \int_0^t \left(M_{ij} \left(\mathbf{z}(t), \mathbf{z}(\tau), \frac{t-\tau}{\epsilon} \right) - M_{ij} \left(\mathbf{z}'(t), \mathbf{z}'(\tau), \frac{t-\tau}{\epsilon} \right) \right) V_j'(\tau)\, d\tau. \end{aligned}$$

We may now apply Lemmas 3.7 and 3.8 for sufficiently small $t > 0$. We thus obtain

$$\begin{aligned} |V_i - V_i'|(t) \le {} & C|\mathbf{z} - \mathbf{z}'|(t) \\ & + C \left(\int_0^t |V_i - V_i'|(\tau)\, d\tau + \int_0^t (t-\tau)^{-1/2} (|\mathbf{z}-\mathbf{z}'|(t) + |\mathbf{z}-\mathbf{z}'|(\tau))\, d\tau \right). \end{aligned}$$

But $|z_i - z_i'|(t) = |\int_0^t (V_i - V_i')(\tau)\, d\tau|$, so having this in mind we arrive at

$$|\mathbf{V} - \mathbf{V}'|(t) \le C \left(2(1+\sqrt{t}) \int_0^t |\mathbf{V} - \mathbf{V}'|(\tau)\, d\tau + \int_0^t (t-\tau)^{-1/2} \int_0^\tau |\mathbf{V} - \mathbf{V}'|(\sigma)\, d\sigma \right).$$

Changing the order of integration in the double integral yields

$$|\mathbf{V} - \mathbf{V}'|(t) \le C' \int_0^t |\mathbf{V} - \mathbf{V}'|(\tau)\, d\tau.$$

Now, by Grönwall's inequality we obtain that

$$\mathbf{V}(\tau) = \mathbf{V}'(\tau)$$

for all $\tau \in [0, t]$, as desired. □

We remark that the above representation of $\mathbf{V}$ does not yield an improved regularity solution.

4. Geometric estimates. We gather here some estimates which are necessary for the next section but they may be of independent interest. We first state the underlying assumption for the rest of the paper, namely,

(W) the Wulff shape W is a regular polygon with N sides, and the distance from the center of symmetry to its facets is $d > 0$.

It is clear that if we fix the inner angle of a polygon, then if the quantity

$$\frac{\max_{i=1,\dots,N} L_i}{\min_{i=1,\dots,N} L_i} \tag{4.1}$$

is bounded, then the isoperimetric ratio L^2/A is finite. But in general the converse is not true. One can devise a sequence of polygons $\{\gamma_n\}_{n=1}^{\infty}$ for which the quotients $(L^n)^2/A^n$ remain bounded but (4.1) explodes. On the other hand we depend on boundedness of (4.1); see the next section or [Ry1, Theorem 10]. Here we prove that the needed estimate holds but only for polygons with L^2/A only slightly bigger than for the regular polygon.

We also recall the estimates for the extinction time for regular polygons moving by crystalline curvature

$$V_i = \frac{\Gamma_i}{\beta_i L_i}. \tag{4.2}$$

Such a bound combined with the comparison principle of Giga and Gurtin [GG] will provide estimates for maximal existence times of solution to (4.2) for any initial polygon.

We now introduce some simplifying notation. We also recall that $\mathcal{S}$ is the set of outer normals to W. Let us suppose that γ is a convex polygon. We denote by D_γ the region bounded by γ, $|D_\gamma|$ is its Lebesgue measure, $n(\gamma)$ is the number of (nonzero) facets, and $|\gamma|$ is the perimeter of γ; $\theta = 2\pi/N$. We set $Q(\gamma)$ to be the isoperimetric quotient of γ, i.e.,

$$Q(\gamma) = \frac{|\gamma|^2}{|D_\gamma|}.$$

We restrict our attention to polygons which may be obtained from W by moving its sides in normal directions, where we do not exclude the possibility that some of the facets get lost. By definition, $\mathcal{P}$ is the set of all planar polygons γ such that

(a) D_γ is convex;

(b) $n(\gamma) \le N$;

(c) the outer normals to the edges to γ belong to $\mathcal{S}$.

We will show this in the following theorem.

THEOREM 4.1. *There exist positive numbers $\lambda_1 > Q(W)$ and Λ_1 such that if $\gamma \in \mathcal{P}$ and $Q(\gamma) \le \lambda_1$, then*

$$\frac{\max_{i=1,\dots,N} L_i(\gamma)}{\min_{i=1,\dots,N} L_i(\gamma)} \le \Lambda_1.$$

Remark. This theorem is trivial for $N = 4$; we leave to the reader to work out the formula expressing $\max\{a, b\}/\min\{a, b\}$ in terms of the isoperimetric quotient of rectangle with sides a and b. We will consider only $N \geq 5$. In this case the bound on λ_1 is constructive while that for Λ_1 is not.

The proof of this theorem requires a lemma.

LEMMA 4.2. *There exists $\alpha > 0$ such that for any $\gamma \in \mathcal{P}$ having less than N sides the following inequality holds:*

$$Q(\gamma) \geq \omega_N + \alpha,$$

where $\omega_N := \frac{1}{4}Q(W) = N \tan \frac{\pi}{N}$.

Proof. In what follows we use the notation convention of section 2. Let us set $s_0 = W$. Thus, by the very definition of $\mathcal{P}$, if $\gamma \in \mathcal{P}$, then $\gamma = s(\mathbf{z})$, where $\mathbf{z} \in \mathbb{R}^N$ is defined as in (2.1) (to be precise we have to substitute the line l_i containing the ith facet of γ for $l_i(t)$). We recall that Lemma 2.1 guarantees that the mapping $\mathcal{P} \ni \gamma \mapsto Q(\gamma)$ is continuous, because $\gamma = s(\mathbf{z})$.

Our argument is organized in several steps. We claim that the general case may be reduced to $n(\gamma) = N - 1$.

(a) We will first show that if $n(\gamma) < N$, then there exists nearby $\gamma' \in \mathcal{P}$ such that $n(\gamma') = n(\gamma) + 1$ and $Q(\gamma') < Q(\gamma)$. Let us suppose that $n(\gamma) = N - m$, $m \geq 1$, and at vertex v_i the edges $s_{i+1}, \ldots, s_{i+j-1}$ are missing, i.e., $\{v_{i+1}\} = s_i \cap s_{i+j}$, $j > 1$, and $m \geq j - 1$. The angle between the normal to s_i and the normal to s_{i+j} is $j\theta$. We will construct a new polygon γ' by moving facet s_{i+1} of zero length by h into the direction of ν_{i+1} (hence, $\Delta\mathbf{z} = h\mathbf{e}_i$), where $h < 0$, i.e., we move it inward. We also assume that $|h|$ and $|h|/L$ are small, where $L = |\gamma|$ is the perimeter of γ. We note that by the definition of $\mathcal{P}$ the angle between ν_i and ν_{i+1} is θ and the angle between ν_{i+1} and ν_{i+j} is $k\theta$, where $k = j - 1 > 0$. By Lemma 2.1 we have

$$\begin{aligned} \Delta L_i &= h/\sin\theta, \\ \Delta L_{i+1} &= -h(\operatorname{ctan}\theta + \operatorname{ctan}k\theta), \\ \Delta L_{i+j} &= h/\sin k\theta; \end{aligned}$$

then

$$\Delta L = \left(\tan\frac{\theta}{2} + \tan k\frac{\theta}{2}\right) h < 0.$$

The change in the area ΔA is also negative and

$$\Delta A = \frac{1}{2}h|\Delta L_i| = -\frac{1}{2}h^2(\operatorname{ctan}\theta + \operatorname{ctan}k\theta) =: -h^2 c(\theta).$$

Now,

$$Q(\gamma') = \frac{(L + \Delta L)^2}{A + \Delta A}.$$

We note that Taylor's expansion yields

$$\begin{aligned} \frac{1}{A + \Delta A} &= \frac{1}{A} \cdot \frac{1}{1 + \Delta A/A} = \frac{1}{A}\left(1 - \frac{\Delta A}{A} + \left(\frac{\Delta A}{A}\right)^2 \int_0^1 \frac{2(1-s)}{(1 + s\Delta A/A)^3}\, ds\right) \\ &\leq \frac{1}{A}\left(1 - \frac{\Delta A}{A} + 8\left(\frac{\Delta A}{A}\right)^2\right), \end{aligned}$$

provided that

$$\left|\frac{h}{L}\right| \le \sqrt{\frac{2A}{L^2 c(\theta)}}. \tag{4.3}$$

By simple algebra we arrive at

$$Q(\gamma') \le \frac{L^2}{A}\left(1+\frac{4h}{L}\left(\tan\frac{\theta}{2}+\tan k\frac{\theta}{2}\right)+\left(\frac{h}{L}\right)^2\left(4\left(\tan\frac{\theta}{2}+\tan k\frac{\theta}{2}\right)^2+c(\theta)\frac{L^2}{A}\right)\right),$$

provided that

$$\left|\frac{h}{L}\right| \le \frac{\tan\frac{\theta}{2}+\tan k\frac{\theta}{2}}{(\tan\frac{\theta}{2}+\tan k\frac{\theta}{2})^2+2c(\theta)Q(\gamma)} \tag{4.4}$$

and

$$\left|\frac{h}{L}\right| \le \frac{1}{\tan\frac{\theta}{2}+\tan k\frac{\theta}{2}}. \tag{4.5}$$

We note that (4.4) is more restrictive than (4.3) and (4.5). But neither of them takes into account the length of neighboring facets s_i, s_{i+j}. We have to restrict the size of $|h|$ again in order to guarantee that $n(\gamma') = n(\gamma)+1$. Finally,

$$Q(\gamma') \le Q(\gamma)\left(1+\frac{2h}{L}\left(\tan\frac{\theta}{2}+\tan k\frac{\theta}{2}\right)\right) < Q(\gamma),$$

provided that h/L is sufficiently small.

(b) If $n(\gamma) = N-m$, $m>1$, then by step (a) we may construct polygons γ_k, $k=1,\dots,m-1$, which are all close to γ and

$$n(\gamma_k) = N-m+k, \quad Q(\gamma_{k-1}) > Q(\gamma_k), \qquad k=1,\dots,m-1,$$

where $\gamma_0 := \gamma$. It follows that

$$Q(\gamma) > Q(\gamma_{m-1}), \qquad n(\gamma_{m-1}) = N-1.$$

(c) The general case has been now reduced to $n(\gamma) = N-1$. Obviously we have

$$Q(\gamma) \ge \min\{Q(\gamma_1):\ \gamma_1\in\mathcal{P},\ n(\gamma_1)=N-1\}.$$

We claim that this minimum is attained. As a matter of fact this follows from the Hausdorff selection theorem but we prefer to prove it directly. Let us suppose that $\{\gamma^n\}_{n=1}^\infty \subset \mathcal{P}$ is a minimizing sequence. After rescaling we may assume that $|D_{\gamma^n}| = 1$. Then,

$$|s(\mathbf{z}^n)|^2 \le 1+\min\{Q(\gamma_1):\ \gamma_1\in\mathcal{P},\ n(\gamma_1)=N-1\} =: A$$

for sufficiently large n. After shifting the polygons we may assume that they all lay in a ball $B(0,A^{1/2})$. Hence the sequence $\{\mathbf{z}^n\}_{n=1}^\infty \subset \mathbb{R}^N$ defining them (i.e., $\gamma^n = s(\mathbf{z}^n)$) is bounded. We may extract a convergent subsequence, still denoted $\{\mathbf{z}^n\}_{n=1}^\infty$, and $\mathbf{z}^n \to \mathbf{z}^\infty$. It follows from Lemma 2.1 that

$$Q(\gamma^n) \to Q(s(\mathbf{z}^\infty)) = \min\{Q(\gamma_1):\ \gamma_1\in\mathcal{P},\ n(\gamma_1)=N-1\}.$$

(d) By the argument in (a) there exists a polygon $\gamma' \in \mathcal{P}$, which is close to $s(\mathbf{z}^\infty)$, $n(\gamma') = N$ and which satisfies

$$Q(\gamma') < Q(s(\mathbf{z}^\infty)) \le Q(\gamma).$$

Of course,

$$Q(\gamma') \ge \min\{Q(\gamma_1) :\ \gamma_1 \in \mathcal{P},\ n(\gamma_1) = N\} = Q(W).$$

The compactness argument in (c) shows that the minimum is attained. For the sake of completeness we should show that indeed the minimum is equal to $Q(W)$. As a matter of fact a more general result is true. Stancu proved (see [St, Theorem 4.2]) that for any $\gamma \in \mathcal{P}$ with N sides the following inequality holds:

$$B_W(\rho) := \rho|\gamma| - |D_\gamma| - \omega_N \rho^2 \ge 0 \tag{4.6}$$

for all $\rho \in [\rho_{in}, \rho_{out}]$. For the moment the specific definitions of $0 < \rho_{in} < \rho_{out}$ are not important. We will present them later. We note that (4.6) is equivalent to

$$\left(\omega_N \rho - \frac{|\gamma|}{2\sqrt{\omega_N}}\right)^2 \le \frac{|\gamma|^2 - 4\omega_N |D_\gamma|}{4\omega_N},$$

hence

$$Q(\gamma) = \frac{|\gamma|^2}{|D_\gamma|} \ge 4\omega_N = Q(W).$$

Combining these estimates we conclude that

$$Q(\gamma) \ge \min_{\{\gamma_1 \in \mathcal{P}: n(\gamma_1) = N-1\}} Q(\gamma_1) = Q(W) + \alpha,$$

where

$$\alpha = \min_{\{\gamma \in \mathcal{P}: n(\gamma) = N-1\}} Q(\gamma) - \min_{\{\gamma \in \mathcal{P}: n(\gamma) = N\}} Q(\gamma) > 0. \qquad \square$$

We are now ready for the proof.

Proof of Theorem 4.1. We take $\lambda_1 = \omega_N + \alpha/2$, but let us suppose that the desired number Λ_1 does not exist, i.e., there is a sequence of polygons $\{\gamma^n\}_{n=1}^\infty \subset \mathcal{P}$ such that

$$\frac{\max L_i^n}{\min L_i^n} \to \infty \quad \text{as } n \to \infty, \tag{4.7}$$

and

$$Q(\gamma^n) \le \lambda_1.$$

All the polygons γ^n, $n = 1, 2, \ldots$, have N sides, so there exists at least one pair of indices (l, j) such that $\max_k L_k^n = L_l^n$ and $\min_k L_k^n = L_j^n$ for infinitely many n's. After choosing an appropriate subsequence of $\{\gamma^n\}_{n=1}^\infty$, which is denoted again by $\{\gamma^n\}_{n=1}^\infty$ condition (4.7) implies that

$$L_l^n / L_j^n \to \infty$$

as n goes to infinity.

Possibly after scaling the polygons, we may assume that $|D_{\gamma^n}| = 1$. Subsequently, we apply the compactness argument as in step (c) in the proof of Lemma 4.2. Hence, we may assume $\mathbf{z}^n \to \mathbf{z}^\infty$, where $\gamma^n = s(\mathbf{z}^n)$. In particular we have $L_i^n \to L_i^\infty$, $i = 1, \dots, N$.

By (4.7) for any $\delta > 0$ there exists n_0 such that

$$L_l^n / L_j^n > 1/\delta$$

for $n \geq n_0$, or

$$\frac{1}{2}\delta\sqrt{\lambda_1} \geq \delta L_l^n > L_j^n.$$

This implies that L_j^n converges to zero. But, γ^∞ is not a point since the area of D_{γ^∞} is one. Hence, $n(\gamma^\infty) < N$ and this implies that

$$\lambda_1 \geq Q(\gamma^n) \to Q(\gamma^\infty) \geq \omega_N + \alpha > \lambda_1.$$

We reached a contradiction, and our theorem follows. □

We now gather estimates for solutions of (4.2). We start with a definition

$$\rho_{out} = \inf\{\mu : \exists p \in \mathbb{R}^2,\ \mu W + p \supset D_\gamma\},$$
$$\rho_{in} = \sup\{\mu : \exists p \in \mathbb{R}^2,\ \mu W + p \subset D_\gamma\}.$$

The next lemma applies to admissible polygons. Let us recall that admissibility was defined at the beginning of section 2.

LEMMA 4.3. *Suppose γ is an admissible convex polygon, then*

(a) $$\rho_{out} \leq L + \frac{\sqrt{L^2 - 4\omega_N A}}{2\omega_N},$$

(b) $$2A/L \geq \rho_{in} \geq A/L.$$

Proof. (a) Stancu proved (see [St, Theorem 4.2]) that

$$B_W(\rho) := \rho L - A - \omega_N \rho^2 \geq 0$$

for all $\rho \in [\rho_{in}, \rho_{out}]$. Thus solving the equation $B_W(\rho) = 0$ provides an upper bound for ρ_{out}.

(b) We note

$$A = \frac{1}{2}\sum_{i=1}^{N} d_i L_i \geq \frac{1}{2}\sum_{i=1}^{N} \rho_{in} L_i = \frac{1}{2}\rho_{in} L.$$

On the other hand $B_W(\rho_{in}) \geq 0$ implies that

$$\rho_{in} L \geq A + \omega_N \rho_{in}^2 \geq A.$$

Part (b) follows. □

This lemma tells us how much we have to scale the Wulff shape so that it contains (or is contained) after scaling and translation a given (respectively, in a given) admissible convex polygon. We find it useful for the estimating of the times of extinction for solutions of (4.2). We recall the lemma below.

LEMMA 4.4 (see [Ry2, Lemma 4.3]). *Let us assume that* $s_0 = \sigma W + p$ *and it evolves according to* (4.2)*; then*

$$L(t) = c_w(T_{\max} - t)^{1/2},$$

where

$$c_w = N\left(\frac{2\kappa\Gamma}{\beta}\right)^{1/2}, \qquad T_{\max} = \left(\frac{L(0)}{c_w}\right)^2 = \frac{L^2(0)\beta}{2N^2\kappa\Gamma}. \qquad \square$$

This lemma gives also the estimates for extinction time of the flow (4.2) for arbitrary convex initial polygon s_0. If s_0 is given then Lemma 4.3 yields $\bar{\gamma}$, $\underline{\gamma}$ such that

$$D_{\bar{\gamma}} \supset D_{s_0} \supset D_{\underline{\gamma}},$$

where $\bar{\gamma} = \rho_{out}W + p_1$, $\underline{\gamma} = \rho_{in}W + p_2$. The first comparison principle of Giga and Gurtin [GG, Section 4] implies that

$$D_{\bar{\gamma}(t)} \supset D_{s(t)} \supset D_{\underline{\gamma}(t)} \tag{4.8}$$

if all polygons evolve according to (4.2), for as long as the solutions exist. Let us denote by $T(s_0)$ the maximal time of existence of solution to (4.2) for a given s_0. We are now in a position to estimate $T(s_0)$ in terms of s_0. The starting point is inclusion (4.8), which shows that

$$T(\underline{\gamma}) \le T(s_0) \le T(\bar{\gamma}).$$

Lemma 4.4 permits us to estimate $T(\bar{\gamma})$ as well as $T(\underline{\gamma})$. We carry out the calculations for $T(\bar{\gamma})$. It is easy to find out $|\bar{\gamma}|$ knowing ρ_{out}; hence by definition of Γ_i's, (2.4) and (2.5) we come to

$$T(\bar{\gamma}) \le \frac{4\beta N^2 \tan^2\frac{\theta}{2}\rho_{out}^2}{2N^2\kappa\Gamma} = \frac{\beta}{2d}\rho_{out}^2.$$

Lemma 4.3 now yields

$$T(\bar{\gamma}) \le \frac{\beta|s_0|^2}{2d}\left(\frac{1+\sqrt{1-Q(W)/Q(s_0)}}{\omega_N}\right)^2. \tag{4.9}$$

Similar calculations lead us to

$$T(\underline{\gamma}) = \frac{\beta}{2d}\rho_{in}^2 \ge \frac{\beta|s_0|^2}{2dQ^2(s_0)}. \tag{4.10}$$

We finally recall the following lemma.

LEMMA 4.5. *If* $V_i < 0$, κ *is defined by* (2.4)*, then*

$$L' = -\sum_{i=1}^{N} V_i\kappa < 0.$$

This follows immediately from Lemma 2.1. $\square$

5. Properties of solutions if the perimeter of s_0 is small. We would like to exhibit in this section some geometric properties of weak solutions in case of convex s_0. We keep the basic assumption (W) of the previous section. Thus, it follows immediately that $\Gamma_i = \Gamma < 0$, $i = 1, \dots, N$. We also assume that all β_i are equal to $\beta > 0$.

We first show that if perimeter L_0 is sufficiently small, and $Q(s_0)$ is close to $Q(W)$ and $u_0 \leq 0$ then the interface $s(t)$ shrinks to a point and $Q(s(t))$ is a decreasing function of time, moreover the temperature remains negative. Of course L_0 must be in some balance with the size of initial distribution of temperature u_0. We may say that for small s_0 the surface tension prevails over the destabilizing bulk forces.

Let us point out a possible interpretation of our result. Since during the evolution the isoperimetric quotient decreases, it seems there is no need for creating new facets during the evolution provided L_0 is already small, as this would increase the isoperimetric quotient. A similar point of view on the problem of whether or not to split a facet is provided by M.-H. Giga and Y. Giga in [Gi]. Their setting, however, is different from ours; they consider curve moving by crystalline curvature under the driving force. Their results are not directly applicable. Nonetheless, we would like to point to [Gi, Lemma 5]. This lemma states that small facets of polygons evolving by crystalline curvature do not split. Obviously, this does not settle the matter here. More work should be done.

After making these remarks let us return to the topic of this section. The main method we use here is based on a priori estimates for

$$\sum_{i=1}^{N} \left| \int_{s_i} u \, dl \right|,$$

and it aims at showing that the time derivative of $Q(s(t))$ is nonpositive. This is the place where the estimates of section 4 come into play. We compare $s(t)$ to evolution of s_0 under the flow of

$$V_i = \frac{\Gamma}{\beta' L_i}, \tag{5.1}$$

where $\beta' > 0$ is chosen appropriately. Here we exploit the comparison principle of Giga and Gurtin [GG].

The last result of this section is concerned with behavior of temperature u as $t \to T_max$. Because at that instance facets move very fast it is not clear whether or not u blows up. It turns out that we are able to show a bound in the $L^\infty(\Omega)$ norm for u in terms of initial data.

We also note a by-product of Theorem 4.1. Namely, we are able to improve Theorem 10 of [Ry1]; we can now show the corollary.

COROLLARY 5.1. *If $(\mathbf{z}, u)$ is a unique weak solution of the quasi-steady approximation of* (2.7)*, i.e., $(\mathbf{z}, u)$ is a solution of*

$$0 = -\int_\Omega \nabla u(x) \cdot \nabla h(x) dx + \sum_{j=1}^{N} \int_{s_j(t)} V_j(t) h(x) dl \qquad \text{for all} \quad h \in H_0^1(\Omega),$$

$$\int_{s_j(t)} u \, dl = \Gamma_j - \beta_j L_j(t) V_j(t), \quad j = 1, \dots, N, \tag{5.2}$$

where s_0 is convex and such that

(a) L_0 *is sufficiently small* (*see Theorem* 10 *in* [Ry1]*)*,
(b) $Q(s_0) \le \lambda_1$,
then

$$\frac{d}{dt} Q(s(t)) \le 0.$$

Proof. We proved in [Ry1, Theorem 10] that $Q(s(t))' \le 0$ provided that

$$\max L_i / \min L_i \le \Lambda_1$$

remains bounded independent of time. Theorem 4.1 provides such a bound if $Q(s_0) \le \lambda_1$. Thus the set $\mathcal{P} \cap \{\gamma : \ |\gamma| \text{ is small, } Q(\gamma) \le \lambda_1\}$ is invariant under the flow of (5.2) and the corollary follows. □

Suppose now we are given an arbitrary convex, an admissible polygon s_0, and a number δ, $\delta \in (0,1)$. Let us now define $\bar{s}(t)$ (respectively, $\underline{s}(t)$) as unique solutions to (5.1) with $\beta' = \beta/(1-\delta)$, (respectively, $\beta' = \beta/(1+\delta)$) and $\underline{s}(0) = s_0$ (respectively, $\bar{s}(0) = s_0$). By (4.9) we have the estimates for T_U the extinction time of $\bar{s}(t)$

$$T_U = T(\bar{\gamma}) \le \frac{\beta |s_0|^2}{2(1-\delta)d} \left(\frac{1 + \sqrt{1 - Q(W)/Q(s_0)}}{\omega_N} \right)^2. \tag{5.3a}$$

Similarly, we have a bound on T_L the extinction time of $\underline{s}(t)$. The formula (4.10) yields

$$T_L = T(\underline{\gamma}) \ge \frac{\beta |s_0|^2}{2(1+\delta) d Q^2(s_0)}. \tag{5.3b}$$

We will compare the flows generated by (5.1) and (2.7), and from the properties of solutions to (5.1) we will infer the behavior of solutions to (2.7). For instance T_U and T_L defined by (5.3) provide upper and lower estimates for T_max.

Here is our main result. It is a strengthened version of [Ry2, Theorem 4.1] which was established only for s_0 being a scaled Wulff shape.

THEOREM 5.2. *Let us suppose that* $0 < \delta < 1$ *is fixed,* s_0 *is a given convex, admissible polygon, and* u_0 *satisfies the condition* (3.4). *The numbers* λ_1 *and* Λ_1 *are given by Theorem* 4.1. *We assume that* $(\mathbf{z}, u)$ *is the unique weak solution with initial conditions* s_0 *and* u_0. *Let us finally suppose that the data fulfill*

(i) $u_0 \le 0$;
(ii) $Q(s_0) \le \lambda_1$;
(iii) $\sum_{i=1}^N |\int_{s_{0i}} u_0 \, dl| \le \exp(C_1 T_U)(C_R L_0^{1/2} \|u_0\|_{H_0^1(\Omega)} + C_2 T_U) \le \delta |\Gamma|$,
where C_R *is defined in* (5.9) *below,*

$$C_1 = N^{3/2} \Lambda_1^{1/2} \beta^{-1} C_0, \quad C_2 = N^{5/2} \Lambda_1^{1/2} |\Gamma| \beta^{-1} C_0,$$

T_L *and* T_U *are given by* (5.3), *and* C_0 *is defined in* (5.7) *below. Then the solution* $(\mathbf{z}, u)$ *satisfies*

(a) $\sum_{i=1}^N |\int_{s_i(t)} u \, dl| \le \exp(C_1 t)(C_R L_0^{1/2} \|u_0\|_{H_0^1(\Omega)} + C_2 t)$, $t \in [0, T_{\max})$;
(b) $V_i(t) < 0$, $i = 1, \dots, N$ *for all* $t \in (0, T_{\max})$, $T_L \le T_{\max} \le T_U$;
(c) $u(t,x) < 0$ *for all* $t \in (0, T_{\max})$, $x \in \Omega$;
(d) $Q(s(t))$ *is a decreasing function of time*;

(e) $s(t)$ *shrinks to a point as* $t \to T_{\max}$.

Proof. We first show that if (a) holds for $t \in [0,\tau]$, where $\tau \in (0, T_max)$, then (b)–(d) are satisfied on this interval.

We begin with proving (b). If (a) holds, then

$$\left| \int_{s_i(t)} u\, dl \right| \le \sum_{i=1}^{N} \left| \int_{s_i(t)} u\, dl \right| \le \delta |\Gamma|$$

and

$$\frac{(1+\delta)\Gamma}{\beta_i L_i} < V_i < \frac{(1-\delta)\Gamma}{\beta_i L_i} < 0, \quad i = 1, \dots, N. \tag{5.4}$$

We have already defined $\bar{s}(t)$ and $\underline{s}(t)$. We immediately obtain from the second comparison theorem of Giga and Gurtin [GG, Section 4] that $D_{\underline{s}(t)} \subset D_{\bar{s}(t)}$ for all t such that $\underline{s}(t)$ is defined. We would be able to estimate T_max above and below only after establishing that

$$\begin{aligned} &D_{\underline{s}(t)} \subset D_{s(t)} \qquad \text{for } t \le \min\{\tau, T_L\}, \\ &D_{s(t)} \subset D_{\bar{s}(t)} \qquad \text{for } t \le \min\{\tau, T_U\}. \end{aligned} \tag{5.5}$$

We cannot use the comparison principles of [GG] because they apply only to solutions of an ODE generalizing (5.1). We will show (5.5) directly.

Let us set

$$E = \{t \in [0,\tau] :\ D_{s(\zeta)} \subset D_{\bar{s}(\zeta)} \text{ for all } \zeta \in [0,t]\};$$

of course $E \neq \emptyset$ since $0 \in E$. We now set $t_1 = \sup E$. We claim that $t_1 = \tau$. Let us suppose the contrary, i.e., $t_1 < \tau$. Hence, by the very definition of t_1 as well as continuity of motion of $s(\zeta)$ and $\bar{s}(\zeta)$ we infer that $s(t_1)$ must touch $\bar{s}(t_1)$. It follows that for some $i \in \{1, \dots, N\}$

$$s_i(t_1) \cap \bar{s}_i(t_1) \neq \emptyset.$$

Let us denote by $\mathcal{I}$ the set of i with the above property, and $\mathcal{I}' = \{1, \dots, N\} \setminus \mathcal{I}$. If $i \in \mathcal{I}$, then because of $D_{s(t_1)} \subset D_{\bar{s}(t_1)}$ we have $|s_i(t_1)| \le |\bar{s}_i(t_1)|$. Inequality (5.4) now yields

$$V_i(t_1) < \frac{(1-\delta)\Gamma}{\beta L_i(t_1)} = \frac{(1-\delta)\Gamma}{|s_i(t_1)|} \le \frac{(1-\delta)\Gamma}{|\bar{s}_i(t_1)|} < 0.$$

We thus conclude that for $i \in \mathcal{I}$ $s_i(t)$ moves inward faster than $\bar{s}_i(t)$ for $t \in [t_1, t_1+\eta)$, η. On the other hand, if $i \in \mathcal{I}'$, then $s_i(t_1) \cap \bar{s}_i(t_1) = \emptyset$. Hence by the continuity of motion, $s_i(t)$ and $\bar{s}_i(t)$ will be separated for $t \in [t_1, t_1+\eta_1)$, $\eta_1 > 0$. We now conclude that $D_{s(\zeta)} \subset D_{\bar{s}(\zeta)}$ for $\zeta \in [t_1, t_1 + \min\{\eta, \eta_1\})$. This contradicts the definition of t_1; hence the second inclusion of (5.5) follows and

$$\tau \le T_U.$$

A similar argument proves the first inclusion of (5.5).

In order to prove that (a) implies (c) we use the variation of constants formula (3.6) and $u = -\Delta U$, and we obtain

$$u(t) = e^{\Delta t/\epsilon} u_0 - \frac{1}{\epsilon} \sum_{i=1}^{N} \int_0^t \Delta e^{\Delta(t-\sigma)/\epsilon} f_i(\mathbf{z}(\sigma)) V_i(\sigma)\, d\sigma.$$

We showed in [Ry2, Lemma 4.5] that $\Delta e^{\Delta t/\epsilon} f_i \le 0$ for $t > 0$, so since $V_i < 0$ we infer that the above integral is nonpositive. By the maximum principle $e^{\Delta t/\epsilon} u_0 < 0$, and (c) follows.

(d) We may now calculate the derivative of the isoperimetric quotient $Q(s(t))$. Since $u(t) < 0$ for $t \in (0, \tau)$ we apply the reasoning as in the proof of Theorem 10 in [Ry1], and we come to

$$\frac{d}{dt} Q(s(t)) \le -\frac{L}{A\beta} \left(\sum_{j=1}^{N} \frac{\kappa^2}{L_j} + \frac{1}{2} \frac{L}{A} N\kappa \right) \left(d - \frac{1}{\kappa N} \sum_{j=1}^{N} \int_{s_j} u \, dl \right),$$

where $d = |\Gamma/\kappa|$ (see the definition of Γ_i's and (2.5)). By Lemmas 11 and 12 of [Ry1] it follows that

$$\sum_{j=1}^{N} \frac{\kappa^2}{L_j} + \frac{1}{2} \frac{L}{A} N\kappa \ge 0.$$

On the other hand since (a) and (b) hold, then

$$d - \frac{1}{N\kappa} \int_{s(t)} u \, dl = \frac{1}{N\kappa} \sum_{i=1}^{N} \left(\Gamma - \int_{s_i(t)} u \, dl \right) \ge 0.$$

Thus,

$$\frac{d}{dt} Q(s(t)) \le 0 \quad \text{on } [0, \tau)$$

(see also proof of Theorem 4.1 in [Ry2]).

We will prove now that (a) holds for $t \in [0, T_max)$. Let us set

$$E = \left\{ t \in [0, T_{\max}) : \sum_{i=1}^{N} \left| \int_{s_i(\tau)} u \, dl \right| \le \exp(C_1 \tau)(C_R L_0^{1/2} \|u_0\|_{H_0^1(\Omega)} + C_2 \tau) \forall \tau \in [0, t) \right\}.$$

Of course $E \ne \emptyset$ because our assumptions imply that $0 \in E$. Let us set $\omega = \sup E$. We shall show that $\omega = T_max$. Let us suppose that $\omega < T_max$; then by definition of ω and (iii)

$$\sum_{i=1}^{N} \left| \int_{s_i(\omega)} u \, dl \right| \le \exp(C_1 \omega)(C_R L_0^{1/2} \|u_0\|_{H_0^1(\Omega)} + C_2 \omega) < \delta |\Gamma|.$$

Thus, there exists $\eta > 0$ such that for $\tau \in [\omega, \omega + \eta)$ we have

$$\sum_{i=1}^{N} \left| \int_{s_i(\tau)} u \, dl \right| < \delta |\Gamma|.$$

Let us note that by (2.7b) and Lemma 3.5, (5.6) takes the form

$$u(t) = e^{\Delta t/\epsilon} u_0 - \frac{1}{\epsilon} \sum_{i=1}^{N} \int_0^t e^{\Delta(t-\tau)/\epsilon} \int_{s_i(\tau)} G(x, y, (t-\tau)/\epsilon) \, dy \frac{\Gamma - \int_{s_i(\tau)} u \, dl}{\beta L_i(\tau)} \, d\tau.$$

Hence,

$$\left|\int_{s_j(t)} u\,dl\right| \le \left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right| + \sum_{i=1}^{N}\left|\int_0^t\int_{s_j(t)}\int_{s_i(\tau)} \frac{1}{\epsilon}G(x,y,(t-\tau)/\epsilon)\,dydl(x)\frac{\Gamma-\int_{s_i(\tau)}u\,dl}{\beta L_i(\tau)}\,d\tau\right|.$$

We note that since all V_i are negative, then $s(t)$ for $t\in[0,\omega)$ are contained in $D_{s_0}$and

$$\int_{s_i(t)}\int_{s_j(\tau)} G^2(x,y,(t-\tau)/\epsilon)\,dydx \le C_0 < \infty, \tag{5.7}$$

where C_0 is independent of time. Hence, the Schwarz inequality implies

$$\begin{aligned}\left|\int_{s_j(t)} u\,dl\right| &\le \left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right| + C_0\beta^{-1}\sum_{i=1}^{N}\int_0^t L_j^{1/2}(t)L_i^{-1/2}(\tau)\left(|\Gamma|+\left|\int_{s_i(\tau)}u\,dl\right|\right)d\tau\\ &\le \left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right| + C_0\beta^{-1}\sum_{i=1}^{N}\int_0^t \frac{L^{1/2}(t)}{\min L_k^{-1/2}(\tau)}\left(|\Gamma|+\left|\int_{s_i(\tau)}u\,dl\right|\right)d\tau.\end{aligned}$$

In order to estimate it further we recall that by Theorem 4.1 $\min L_i \ge \Lambda_1^{-1}\max L_i$ and $N\max L_i \ge L$. Thus, we arrive at

$$\left|\int_{s_j(t)} u\,dl\right| \le \left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right| + C_0\beta^{-1}\Lambda_1^{1/2}N^{1/2}\int_0^t L^{1/2}(t)L^{-1/2}(\tau)\left(N|\Gamma|+\sum_{i=1}^{N}\left|\int_{s_i(\tau)}u\,dl\right|\right)d\tau.$$

By Corollary 4.5 $L(t) < L(\tau)$ if $\tau < t$. Therefore, the summing up of the estimates for $|\int_{s_j(t)} u\,dl|$ yields

$$\begin{aligned}\sum_{j=1}^{N}\left|\int_{s_j(t)} u\,dl\right| &\le \sum_{j=1}^{N}\left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right|\\ &\quad + C_0\beta^{-1}\Lambda_1^{1/2}N^{3/2}\int_0^t\left(N|\Gamma|+\sum_{j=1}^{N}\left|\int_{s_j(\tau)}u\,dl\right|\right)d\tau.\end{aligned} \tag{5.8}$$

We are almost in a position to apply Grönwall's inequality, but before doing so we note that

$$\begin{aligned}\sum_{j=1}^{N}\left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right| &\le \sum_{j=1}^{N} L_j^{1/2}(t)\left\|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\right\|_{L^2(s_j(t))}\\ &\le L^{1/2}(t)N^{1/2}c_r\|e^{\Delta t/\epsilon}u_0\|_{H_0^1(\Omega)}\\ &\le L^{1/2}(0)N^{1/2}c_r\|e^{\Delta t/\epsilon}u_0\|_{H_0^1(\Omega)}\end{aligned}$$

and c_r is the constant in the inequality

$$\|u\|_{L^2(l\cap\Omega)} \le c_r \|u\|_{H_0^1(\Omega)};$$

l is any line in $\mathbb{R}^2$.

We now need to estimate $\|e^{\Delta t/\epsilon}u_0\|_{H_0^1(\Omega)}$. This can be done with the help of (3.5) and the properties of fractional powers of $-\Delta$ (see [He, Sections 1.3 and 1.4]). One can see that

$$\|e^{\Delta t/\epsilon}u_0\|_{H_0^1(\Omega)} \le B\|(-\Delta)^{1/2}e^{\Delta t/\epsilon}u_0\|_{L^2(\Omega)} \le BM\|u_0\|_{X^{1/2}} \le B^2M\|u_0\|_{H_0^1(\Omega)}.$$

Combining these estimates we come to

$$\sum_{j=1}^{N}\left|\int_{s_j(t)} e^{\Delta t/\epsilon}u_0\,dl\right| \le C_R L^{1/2}(0)\|u_0\|_{H_0^1(\Omega)},$$

where

$$C_R = c_r N^{1/2}B^2M. \tag{5.9}$$

By Grönwall's inequality applied to (5.8) we obtain

$$\sum_{j=1}^{N}\left|\int_{s_j(t)} u\,dl\right| \le e^{C_1 t}(C_2 t + C_R L_0^{1/2}\|u_0\|_{L^2(\Omega)}),$$

so (a) holds on $[0, \omega+\eta]$ too, contrary to the maximality of ω.

Now, after we established that (a) holds for all $t < T_max$, the first inclusion of (5.5) implies the lower bound on T_max, and the second inclusion of (5.5) implies the upper bound, i.e.,

$$T_U \ge T_{\max} \ge T_L.$$

(e) Since all the velocities V_i's are negative and $Q(s(t))$ decreases in time, then the only possibility of extinction is that $s(t)$ shrinks to a point. □

We showed that integrals of u over $s(t)$ remain bounded throughout the evolution. We shall prove a stronger result, namely that u itself remains bounded.

THEOREM 5.3. *Under the assumptions of the previous theorem, we have*

$$\|u(t)\|_{L^\infty(\Omega)} \le \|u_0\|_{L^\infty(\Omega)} + C(L_0).$$

Proof. We claim that the compatibility condition (3.4) implies that $\|u_0\|_{L^\infty(\Omega)}$ is finite for

$$u_0 - \sum_{j=1}^{N} f_j(0)V_j(0) = h \in H^2(\Omega)\cap H_0^1(\Omega).$$

On the other hand by [Ry2, Lemma 3.3], $f_i(0)$ may be decomposed in the following way:

$$f_i(0) = \varphi g_i + r_i,$$

where $\varphi \in C_0^\infty(\Omega)$, $g_i \in H^\sigma(\mathbb{R}^2)$ ($\sigma \in (1, 3/2)$ is arbitrary), and $r_i \in H^2(\Omega) \cap H_0^1(\Omega)$. Our claim now follows from the Sobolev embedding theorem.

We use the variation of constants formula (3.6). Next, we apply $-\Delta$ to both sides. After setting $u = -\Delta U$ we obtain

$$u(t) = e^{\Delta t/\epsilon} u_0 - \frac{1}{\epsilon} \int_0^t \Delta e^{\Delta(t-\tau)/\epsilon} \sum_{i=1}^N f_i(\tau) V_i(\tau)\, d\tau.$$

By the maximum principle $\|e^{\Delta t/\epsilon} u_0\|_{L^\infty(\Omega)} \le \|u_0\|_{L^\infty(\Omega)}$. Let us pick $\alpha \in (\frac{1}{2}, 1)$, then by the embedding theorem (see [Hn, Theorem 1.6.1]) $X^\alpha \subset C^{0,\mu}(\Omega)$, where $0 \le \mu < 2\alpha - 1$. Hence we have

$$\begin{aligned}
\|u(t)\|_{L^\infty(\Omega)} &\le \|u_0\|_{L^\infty(\Omega)} + \frac{1}{\epsilon} \left\| \int_0^t \Delta e^{\Delta(t-\tau)/\epsilon} \sum_{i=1}^N f_i(\tau) V_i(\tau)\, d\tau \right\|_{X^\alpha} \\
&\le \|u_0\|_{L^\infty(\Omega)} + \frac{1}{\epsilon} \lim_{\eta \to 0^+} \int_0^{t-\eta} \left\| (-\Delta)^{1+\alpha} e^{\Delta(t-\tau)/\epsilon} \sum_{i=1}^N f_i(\tau) V_i(\tau) \right\|_{L^2(\Omega)} d\tau \\
&\le \|u_0\|_{L^\infty(\Omega)} \\
&+ \frac{1}{\epsilon} \lim_{\eta \to 0^+} \int_0^{t-\eta} \left\| (-\Delta)^{\alpha} e^{\Delta \frac{(t-\tau)}{2\epsilon}} \Delta e^{\Delta \frac{(t-\tau)}{2\epsilon}} \sum_{i=1}^N f_i(\tau) V_i(\tau) \right\|_{L^2(\Omega)} d\tau \\
&= \|u_0\|_{L^\infty(\Omega)} + J.
\end{aligned}$$

We estimate J using Lemma 3.5 and [He, Theorem 1.4.3]. This yields

$$J \le C_\alpha \lim_{\eta \to 0^+} \int_0^{t-\eta} \frac{(2\epsilon)^\alpha}{(t-\tau)^\alpha} \sum_{i=1}^N \left\| \int_{s_i(\tau)} G(x, y, \frac{t-\tau}{2\epsilon})\, dl(y) \right\|_{L^2(\Omega)} |V_i(\tau)|\, d\tau.$$

The problem now is to bound the integral involving the Green function. It is a simple task using the Schwarz inequality, (3.12), and an inequality $t^{-2} \exp(-|\xi|^2/t) \le At^{-1} \exp(-|\xi|^2/(2t))$,

$$\begin{aligned}
\int_\Omega \left(\int_{s_i(\tau)} G(x, y, \zeta)\, dl(y) \right)^2 dx &\le \int_\Omega |s_i(\tau)| \int_{s_i(\tau)} G^2(x, y, \zeta)\, dl(y) dx \\
&\le 2L_i(\tau) \int_{s_i(\tau)} \int_\Omega (K_\zeta^2(x-y) + H^2(x, y, \zeta))\, dx dl(y) \\
&\le 2L_i(\tau) \int_{s_i(\tau)} (A + \|H\|^2_{L^\infty(\Omega \times \Omega_h \times [0, T_{\max}])})\, dl(y) \\
&\le A' L_i^2(\tau),
\end{aligned}$$

where $h = \operatorname{dist}(s_0, \partial\Omega)$, $\zeta = (t-\tau)/(2\epsilon)$, and A' is independent of T_max. Combining the above inequality with (5.4) yields

$$J \le C_\alpha \epsilon^\alpha \int_0^t (t-\tau)^{-\alpha} \sum_{i=1}^N \frac{L_i(\tau)(1+\delta)}{\beta L_i(\tau)}\, d\tau.$$

Finally,

$$\|u(t)\|_{L^\infty(\Omega)} \le \|u_0\|_{L^\infty(\Omega)} + J \le \|u_0\|_{L^\infty(\Omega)} + C' T_U^{(1-\alpha)}.$$

Since T_U can be bounded in terms of $L(0)$ our theorem follows. □

We close by stating a simple estimate for perimeter of $s(t)$.

COROLLARY 5.4. *Under the assumptions of Theorem* 5.2*, we have*

$$L(t) \geq C(T_{\max} - t)^{1/2}.$$

Proof. To see this we differentiate $L(t)$

$$L'(t) = -\sum_{i=1}^{N} V_i \kappa = -\kappa\beta^{-1} \sum_{i=1}^{N} \left(\Gamma - \int_{s_i} u \, dl \right) L_i^{-1}(t) \leq -\Gamma(1-\delta)\kappa\beta^{-1} \sum_{i=1}^{N} L_i^{-1}(t).$$

Since $L_i \geq \min L_i \geq \Lambda_1^{-1} \max L_i \geq \Lambda_1^{-1} N^{-1} L$ we obtain for $C_u = \Gamma(1-\delta)\kappa\Lambda_1 N\beta^{-1}$

$$L'(t) \leq -C_u L^{-1}.$$

Our result follows immediately from this differential inequality. □

Acknowledgments. The author wishes to thank Professor Y. Giga for bringing reference [GG] to the author's attention and for stimulating discussions on the subject of this paper. The author is also indebted to the referees for their remarks which led to the improvement of the text and to the removal of some inaccuracies.

REFERENCES

[AW] F. ALMGREN AND L. WANG, *Mathematical existence of crystal growth with Gibbs–Thomson curvature effects*, J. Geom. Anal., to appear.

[BP] BEN AMAR AND Y. POMEAU, *Growth of faceted needle crystals: theory*, Europhys. Lett., 6 (1988), pp. 609–614.

[CR] X. CHEN AND F. REITICH, *Local existence and uniqueness of solutions to the Stefan problem with surface tension and kinetic undercooling*, J. Math. Anal. Appl., 164 (1992), pp. 350–362.

[Gi] M.-H. GIGA AND Y. GIGA, *A subdifferential interpretation of crystalline motion under nonuniform driving force*, in Proceedings, International Conference on Dynamical Systems and Differential Equations, W. Chen and S. Hu, eds., Vol. I, Southwest Missouri State University, Springfield, MO, 1998, pp. 276–287.

[GG] Y. GIGA AND M.E. GURTIN, *A comparison theorem for crystalline evolution in the plane*, Quart. Appl. Math., 54 (1996), pp. 727–737.

[Gu] M. GURTIN, *Thermomechanics of Evolving Phase Boundaries in the Plane*, Clarendon Press, Oxford, UK, 1993.

[GM] M. GURTIN AND J. MATIAS, *Thermomechanics and the formulation of the Stefan problem for fully faceted interfaces*, Quart. Appl. Math., 53 (1995), pp. 761–782.

[GMl] M. GURTIN AND J. MATIAS, *Notes of lectures given by Gurtin at the IMA*, University of Minnesota, Minneapolis, MN, September 1990.

[Hn] D. HENRY, *Geometric theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, Berlin 1981 (Russian translation, Mir, Moscow 1985).

[Hr] C. HERRING, *Surface tension as a motivation for sintering*, in The Physics of Powder Metallurgy, W.E. Kingston, ed., McGraw-Hill, New York, 1958.

[LSU] O.A. LADYZHENSKAYA, V.A. SOLONNIKOV, AND N.N. URALCEWA, *Linear and Quasi-linear Equations of Parabolic Type*, Transl. of Math. Monogr. 23, AMS, Providence, RI, 1968.

[L] S. LUCKHAUS, *Solutions for the two-phase Stefan problem with the Gibbs–Thomson law for the melting temperature*, European J. Appl. Math., 1 (1990), pp. 101–111.

[Ra] E.V. RADKEVICH, *The Gibbs–Thomson correction and condition for the existence of classical solution of the modified Stefan problem*, Soviet Math. Dokl., 43 (1991), pp. 274–278.

[Ry1] P. RYBKA, *A crystalline motion: uniqueness and geometric properties*, SIAM J. Appl. Math., 57 (1997), pp. 53–72.

[Ry2] P. RYBKA, *Crystalline version of the Stefan problem with Gibbs–Thomson law and kinetic undercooling*, Adv. Differential Equations, 3 (1998), pp. 687–713.

[So] H.M. Soner, *Convergence of the phase-field equations to the Mullins-Sekerka problem with kinetic undercooling*, Arch. Rational Mech. Anal., 131 (1995), pp. 134–197.

[St] A. Stancu, *Uniqueness of self-similar solutions for a crystalline flow,* Indiana Univ. Math. J., 45 (1996), pp. 1157–1174.

[T] J.E. Taylor, *Motion of curves by crystalline curvature, including triple junctions and boundary points*, in Differential Geometry: Partial Differential Equations on Manifolds, Proc. Symp. Pure Math. 54, AMS, Providence, RI, 1993, pp. 417–438.

[Tr] H. Triebel, *Theory of Interpolation, Function Spaces, Differential Operators*, North Holland, Amsterdam, 1978.

[TS] A.N. Tychonov and A.A. Samarskii, *Equations of Mathematical Physics*, Macmillan, New York, 1963.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 787–832

ASYMPTOTICS OF SOLUTIONS TO THE MODIFIED NONLINEAR SCHRÖDINGER EQUATION: SOLITONS ON A NONVANISHING CONTINUOUS BACKGROUND*

A. V. KITAEV† AND A. H. VARTANIAN†

Abstract. Using the matrix Riemann–Hilbert factorization approach for nonlinear evolution systems which take the form of Lax-pair isospectral deformations and whose corresponding Lax operators contain both discrete and continuous spectra, we obtain the leading-order asymptotics as $t\to\pm\infty$ of the solution to the Cauchy problem for the modified nonlinear Schrödinger equation, $i\partial_t u+\frac{1}{2}\partial_x^2 u+|u|^2u+is\partial_x(|u|^2u)=0$, $s\in\mathbb{R}_{>0}$, which is a model for nonlinear pulse propagation in optical fibers in the subpicosecond time scale. Also derived are analogous results for two gauge-equivalent nonlinear evolution equations—in particular, the derivative nonlinear Schrödinger equation $i\partial_t q+\partial_x^2 q-i\partial_x(|q|^2q)=0$. As an application of these asymptotic results, explicit expressions for position and phase shifts of solitons in the presence of the continuous spectrum are calculated.

Key words. asymptotics, Riemann–Hilbert problem, solitons, optical fibers

AMS subject classifications. 35Q15, 35Q55, 58F07, 78A60

PII. S0036141098332019

1. Introduction.

With the current emphasis on the utilization of optical fibers, capable of supporting solitons, as the communication channel in the practical realization and implementation of all-optical (lightwave), ultrahigh-bit-rate, long-distance communication systems using the return-to-zero (RZ) format for generating the optical bit stream, design issues requiring the consideration of several factors, e.g., soliton widths and intersoliton spacings, are intimately related to the study of the fundamental dynamical processes associated with the propagation of high-power ultrashort pulses in optical fibers (at the present stage of technology, these systems can at best still only be called near-soliton(ic)-based) [1, 2, 3]. The standard, classical mathematical model for nonlinear pulse propagation in the picosecond time scale in the anomalous dispersion regime in an isotropic, homogeneous, lossless, nonamplifying, polarization-preserving single-mode optical fiber is the nonlinear Schrödinger equation (NLSE) [1, 2, 3, 4]. However, experiments and theories on the propagation of high-power ultrashort pulses in the subpicosecond-femtosecond time scale in monomode optical fibers have shown that the NLSE is no longer a valid model and that additional nonlinear terms (dispersive and dissipative) and higher-order linear dispersion must be taken into account: in this case, pulse-like propagation is described (in dimensionless and normalized form) by the following nonlinear evolution equation (NLEE) [1, 2, 3]:

$$i\partial_z u+\frac{1}{2}\partial_\tau^2 u+|u|^2u+is\partial_\tau(|u|^2u)=-i\widetilde{\Gamma}u+i\breve{\delta}\partial_\tau^3 u+\frac{\widetilde{\tau}_n}{\widetilde{\tau}_0}u\partial_\tau(|u|^2), \tag{1}$$

where u is the slowly varying amplitude of the complex field envelope, z is the propagation distance along the fiber length, τ is the time measured in a frame of reference

*Received by the editors January 5, 1998; accepted for publication July 24, 1998; published electronically May 7, 1999. This research was supported by the Alexander von Humboldt Foundation and partially supported by the Russian Academy of Sciences.
http://www.siam.org/journals/sima/30-4/33201.html
†Steklov Mathematical Institute, Fontanka 27, St. Petersburg 191011, Russia (kitaev@pdmi.ras.ru, arthur@pdmi.ras.ru).

moving with the pulse at the group velocity, $s \in \mathbb{R}_{>0}$ governs the effects due to the intensity dependence of the group velocity (self-steepening), $\widetilde{\Gamma}$ is the intrinsic fiber loss, $\breve{\delta}$ governs the effects of the third-order linear dispersion, and $\widetilde{\tau}_n/\widetilde{\tau}_0$ governs the soliton self-frequency shift effect.

Since, under typical operating conditions, $\widetilde{\Gamma}$, $\breve{\delta}$, and $\widetilde{\tau}_n/\widetilde{\tau}_0$ are small parameters [1, 2, 3], a strategy to study the solutions of (1), for which the nonlinear effects dominate the higher-order linear dispersive effect, is to set the right-hand side equal to zero, thus obtaining the following NLEE (integrable in the sense of the inverse scattering method (ISM) [5, 6, 7]):

$$i\partial_t u + \frac{1}{2}\partial_x^2 u + |u|^2 u + is\partial_x(|u|^2 u) = 0, \tag{2}$$

which hereafter is called the modified nonlinear Schrödinger equation (MNLSE) (the physical variables z and τ have been mapped isomorphically onto the mathematical t and x variables, which are standard in the ISM context) and to treat (1) as a nonintegrable perturbation of the MNLSE. From the above discussion, it is clear that perturbations of multisoliton solutions of the MNLSE can be very important in the physical context, related to optical fibers [1, 2, 3]. Since practical lasers excite not only the soliton(ic) mode(s) but also an entire continuum of linear-like dispersive (radiative) waves, to have physically meaningful and practically representative results, it is necessary to investigate solutions of the MNLSE under general initial (launching, in the optical fiber literature [1, 2, 3]) conditions, without any artificial restrictions and/or constraints, which have both soliton(ic) and nonsoliton(ic) (continuum) components: it is towards such a solution that the initial pulse launched into an optical fiber is evolving asymptotically [8]. In physical terms, the pulse adjusts its width as it propagates along the optical fiber to evolve into a (multi)soliton, and a part of the pulse energy is shed in the form of dispersive waves in the process: normally, these dispersive waves form a low-level broadband background radiation that accompanies the (multi)soliton [1, 2, 3, 8]. From the physical point of view, therefore, it is important to understand how the continuum and the (near)soliton(s) interact and to be able to derive an explicit functional form for this process. Since (2) is integrable via the ISM, we can use one of the techniques developed in the framework of this approach to solve the aforementioned problem; in particular, in this paper the matrix Riemann–Hilbert (RH) factorization method is applied.

For several soliton-bearing equations, e.g., KdV, Landau–Lifshitz, NLS, sine-Gordon and MKdV, it is known that the dominant ($\mathcal{O}(1)$) asymptotic ($t \to \pm\infty$) effect of the continuous spectra on the multisoliton solutions is a shift in phase and position of their constituent solitons [9, 10, 11, 12]: as will be shown in this paper, an analogous, though analytically more complicated, situation takes place for the MNLSE (the additional complexification occurs due to the nonstandard normalization of the associated RH problem). While the abovementioned works deal only with the leading-order ($\mathcal{O}(1)$) asymptotic term, in this paper, for the MNLSE, not only the leading-order, but the next-to-leading-order ($\mathcal{O}(t^{-1/2})$) term as well is derived; in particular, besides inducing an $\mathcal{O}(1)$ position and phase shift on the multisoliton solution, this $\mathcal{O}(t^{-1/2})$ term represents the evolution of the continuum component (dispersive wavetrain [1, 2, 3]) as well as the nontrivial interaction (overlap) of the soliton and continuum components of the solution. It is worth mentioning that, even though there have been several papers [13, 14] devoted to studying the soliton solutions of the MNLSE, to the best of our knowledge, very little, if anything, was known about its solution(s) for the class of nonreflectionless initial data until very recently [15]. In

the framework of the ISM, an asymptotic analysis of the aforementioned solution for the MNLSE can be divided into two stages: (1) the investigation of the continuum (solitonless) component of the solution [16, 17, 18, 19, 20, 21]; and (2) the inclusion of the (multi)soliton component via the application of a "dressing" procedure [22, 23] to the continuum background. In this paper, the abovementioned asymptotic paradigm is carried out systematically for the MNLSE: the results obtained in this paper are formulated as Theorems 2.1–2.3.

This paper is organized as follows. In section 2, a matrix RH problem for the solution of an NLEE gauge-equivalent to equation (2) is stated, and the results of this paper are summarized as Theorems 2.1–2.3. In section 3, an extended RH problem is formulated and shown to be equivalent to the original one stated in section 2, and as $t \to +\infty$, it is shown that the solution of the extended RH problem converges, modulo exponentially decreasing terms, to the solution of a model RH problem. In section 4, the Beals–Coifman [24, 25] formulation for the solution of an RH problem on an oriented contour is succinctly recapitulated, and the model RH problem is solved asymptotically as $t \to +\infty$ for the Schwartz class of nonreflectionless generic potentials. In section 5, a phase integral which is associated with the nonstandard normalization of the abovementioned RH problem is evaluated asymptotically as $t \to +\infty$. Finally, in section 6 the asymptotic analysis as $t \to -\infty$ is presented.

2. The RH problem and summary of results. In this section, the matrix RH problem is stated, and the main results of the paper are formulated as Theorems 2.1–2.3. Before doing so, however, it is necessary to introduce some notation and definitions which are used throughout the paper.

Notational conventions.

(1) $e_{\alpha\beta}$, $\alpha, \beta \in \{1,2\}$, *denote* 2×2 *matrices with entry* 1 *in* $(\alpha\,\beta)$, $(e_{\alpha\beta})_{ij} := \delta_{\alpha i}\delta_{\beta j}$, $i,j \in \{1,2\}$, *where* δ_{ij} *is the Kronecker delta*;
(2) $\mathrm{I} := e_{11} + e_{22} = \mathrm{diag}(1,1)$ *denotes the* 2×2 *identity matrix*;
(3) $\sigma_3 := e_{11} - e_{22} = \mathrm{diag}(1,-1)$, $\sigma_- := e_{21}$, $\sigma_+ := e_{12}$, *and* $\sigma_1 := \sigma_- + \sigma_+$;
(4) *for a scalar* ϖ *and a* 2×2 *matrix* Υ, $\varpi^{\mathrm{ad}(\sigma_3)}\Upsilon := \varpi^{\sigma_3}\Upsilon\varpi^{-\sigma_3}$;
(5) $\overline{(\bullet)}$ *denotes complex conjugation of* $(\bullet)$;
(6) $M_2(\mathbb{C})$ *denotes the* 2×2 *complex matrix algebra with the following inner product* $((\cdot,\cdot): M_2(\mathbb{C}) \times M_2(\mathbb{C}) \to \mathbb{C})$, $\forall\, a, b \in M_2(\mathbb{C})$, $(a,b) := \mathrm{tr}(\bar{b}a)$, *and* (*for* $a \in M_2(\mathbb{C})$) *the norm on* $M_2(\mathbb{C})$ *is defined as* $|a| := \sqrt{(a,a)}$;
(7) $\mathcal{L}^p(D;M_2(\mathbb{C})) := \{f;\, f: D \to M_2(\mathbb{C}),\, ||f||_{\mathcal{L}^p(D;M_2(\mathbb{C}))} := (\int_D |f(\varrho)|^p |d\varrho|)^{1/p} < \infty, p \in \{1,2\}\}$;
(8) $\mathcal{L}^\infty(D;M_2(\mathbb{C})) := \{g; g: D \to M_2(\mathbb{C}), ||g||_{\mathcal{L}^\infty(D;M_2(\mathbb{C}))} := \max_{1\le i,j\le 2} \sup_{\varrho\in D} |g_{ij}(\varrho)| < \infty\}$;
(9) *for* D *an unbounded domain of* $\mathbb{R}\cup i\mathbb{R}$, *let* $\mathcal{S}(D;\mathbb{C})$ (*resp.*, $\mathcal{S}(D;M_2(\mathbb{C}))$) *denote the Schwartz class on* D, *i.e.*, *the class of smooth* $\mathbb{C}$*-valued* (*resp.*, $M_2(\mathbb{C})$*-valued*) *functions* $f(x): D \to \mathbb{C}$ (*resp.*, $F(x): D \to M_2(\mathbb{C})$) *which together with all derivatives tend to zero faster than any positive power of* $|x|^{-1}$ *as* $|x| \to \infty$.

In this paper, as in [15], along with the MNLSE, the following NLEEs are studied:

$$i\partial_t Q + \partial_x^2 Q + iQ^2\partial_x \overline{Q} + \frac{1}{2}Q|Q|^4 = 0, \tag{3}$$

with initial condition $Q(x,0) \in \mathcal{S}(\mathbb{R};\mathbb{C})$, and the derivative nonlinear Schrödinger equation (DNLSE),

$$i\partial_t q + \partial_x^2 q - i\partial_x(|q|^2 q) = 0, \tag{4}$$

with initial condition $q(x,0)\in\mathcal{S}(\mathbb{R};\mathbb{C})$. To recall the relations between the solutions of these NLEEs, the following propositions are formulated.

PROPOSITION 2.1 (see [26]). *The necessary and sufficient condition for the compatibility of the following system of linear ODEs (the Lax pair) for arbitrary $\lambda\in\mathbb{C}$:*

$$\partial_x\Psi(x,t;\lambda)=U(x,t;\lambda)\Psi(x,t;\lambda),\qquad \partial_t\Psi(x,t;\lambda)=V(x,t;\lambda)\Psi(x,t;\lambda), \tag{5}$$

where

$$U(x,t;\lambda)=-i\lambda^2\sigma_3+\lambda(\overline{Q}\sigma_-+Q\sigma_+)-\frac{i}{2}|Q|^2\sigma_3,$$
$$V(x,t;\lambda)=2\lambda^2U(x,t;\lambda)-i\lambda((\partial_x\overline{Q})\sigma_--(\partial_xQ)\sigma_+)+\left(\frac{i}{4}|Q|^4+\frac{1}{2}(\overline{Q}\partial_xQ-Q\partial_x\overline{Q})\right)\sigma_3,$$

is that $Q(x,t)$ satisfies (3).

Proof. Equation (3) is the Frobenius compatibility condition for system (5). □

PROPOSITION 2.2. *Let $Q(x,t)$ be a solution of* (3). *Then there exists a corresponding solution of system* (5) *such that $\Psi(x,t;0)$ is a diagonal matrix.*

Proof. For given $Q(x,t)$, let $\widehat{\Psi}(x,t;\lambda)$ be a solution of system (5) which exists in accordance with Proposition 2.1. Setting $\lambda=0$ in system (5), one gets that $\widehat{\Psi}(x,t;0)=\exp\{-\frac{i\sigma_3}{2}\int_{x_0}^x|Q(\varrho,t)|^2d\varrho\}\widehat{\mathcal{K}}(\lambda)$, for some $x_0\in\mathbb{R}$ and nondegenerate matrix $\widehat{\mathcal{K}}(\lambda)$ which is independent of x and t. The function $\Psi(x,t;\lambda):=\widehat{\Psi}(x,t;\lambda)(\widehat{\mathcal{K}}(\lambda))^{-1}$ is the solution of system (5) which is diagonal at $\lambda=0$. □

PROPOSITION 2.3 (see [27]). *Let $Q(x,t)$ be a solution of* (3) *and $\Psi(x,t;\lambda)$ the corresponding solution of system* (5) *given in Proposition* 2.2. *Set $\Psi_q(x,t;\lambda):=\Psi^{-1}(x,t;0)\Psi(x,t;\lambda)$. Then*

$$\partial_x\Psi_q(x,t;\lambda)=\mathcal{U}_q(x,t;\lambda)\Psi_q(x,t;\lambda),\qquad \partial_t\Psi_q(x,t;\lambda)=\mathcal{V}_q(x,t;\lambda)\Psi_q(x,t;\lambda), \tag{6}$$

where

$$\mathcal{U}_q(x,t;\lambda)=-i\lambda^2\sigma_3+\lambda(\overline{q}\sigma_-+q\sigma_+), \tag{7}$$

$$\mathcal{V}_q(x,t;\lambda)=\begin{pmatrix}-2i\lambda^4-i\lambda^2|q|^2 & 2\lambda^3q+i\lambda\partial_xq+\lambda|q|^2q\\ 2\lambda^3\overline{q}-i\lambda\partial_x\overline{q}+\lambda|q|^2\overline{q} & 2i\lambda^4+i\lambda^2|q|^2\end{pmatrix}, \tag{8}$$

with $q(x,t)$ defined by

$$q(x,t):=Q(x,t)((\Psi^{-1}(x,t;0))_{11})^2 \tag{9}$$

being the "Kaup–Newell" [28] Lax pair for the DNLSE.

Proof. Differentiating $\Psi_q(x,t;\lambda):=\Psi^{-1}(x,t;0)\Psi(x,t;\lambda)$ with respect to x and t and using the fact that $\Psi(x,t;0)=\exp\{-\frac{i\sigma_3}{2}\int_{x_0}^x|Q(\varrho,t)|^2d\varrho\}$, for some $x_0\in\mathbb{R}$, and $\Psi(x,t;\lambda)$ satisfy system (5) for $\lambda=0$ and $\lambda\in\mathbb{C}\setminus\{0\}$, resp., defining $q(x,t)$ as in (9), one gets that $\Psi_q(x,t;\lambda)$ satisfies system (6), where $\mathcal{U}_q(x,t;\lambda)$ and $\mathcal{V}_q(x,t;\lambda)$ are given by (7) and (8): (4) is the Frobenius compatibility condition for system (6). □

PROPOSITION 2.4. *If $q(x,t)$ is a solution of the DNLSE such that $q(x,0)\in\mathcal{S}(\mathbb{R};\mathbb{C})$, then*

$$u(x,t):=\frac{1}{\sqrt{2s}}\exp\left\{\frac{i}{s}\left(x-\frac{t}{2s}\right)\right\}q\left(\frac{t}{s}-x,\frac{t}{2}\right) \tag{10}$$

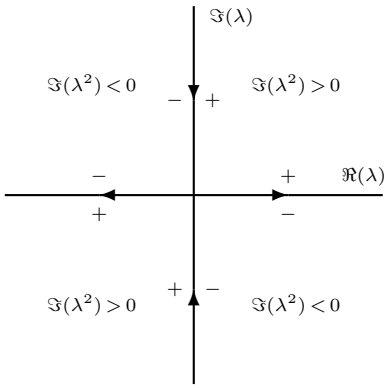


FIG. 1. *Continuous spectrum* $\widehat{\Gamma}$.

satisfies the MNLSE with initial condition $u(x,0)\in\mathcal{S}(\mathbb{R};\mathbb{C})$.

Proof. The proof is by direct substitution. □

Remark 2.1. A convention is now adopted which is adhered to *strictly* throughout the paper: for each segment of an oriented contour, according to the given orientation, the "+" side is to the left and the "−" side is to the right as one traverses the contour in the direction of the orientation; hence, $(\bullet)_+$ and $(\bullet)_-$ denote, resp., the nontangential limits (boundary values) of $(\bullet)$ on an oriented contour from the "+" (left) and "−" (right) sides.

Before stating the matrix RH problem which is investigated asymptotically (as $t\to\pm\infty$) in this paper (see Lemma 2.1), it will be convenient to introduce the following notation: *let* $\mathcal{Z}_d:=\cup_{i=1}^N(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})$ *and* $\widehat{\Gamma}:=\{\lambda;\,\Im(\lambda^2)=0\}$ (*oriented as in Figure* 1) *denote, resp., the discrete and continuous spectra of the operator* $\partial_x-U(x,t;\lambda)$, *and* $\sigma_{\mathcal{L}}:=\mathrm{Spec}(\partial_x-U)=\mathcal{Z}_d\cup\widehat{\Gamma}$ $(\mathcal{Z}_d\cap\widehat{\Gamma}=\emptyset)$.

LEMMA 2.1. *Let* $Q(x,t)$, *as a function of* $x,\in\mathcal{S}(\mathbb{R};\mathbb{C})$. *Set*

$$m(x,t;\lambda):=\Psi(x,t;\lambda)\exp\{i(\lambda^2x+2\lambda^4t)\sigma_3\}.$$

Then (1) *the bounded discrete set* $\mathcal{Z}_d$ *is finite* $(\mathrm{card}(\mathcal{Z}_d)<\infty)$; (2) *the poles of* $m(x,t;\lambda)$ *are simple*; (3) *the first* (*resp., second*) *column of* $m(x,t;\lambda)$ *has poles at* $\{\pm\lambda_i\}_{i=1}^N$ (*resp.,* $\{\pm\overline{\lambda_i}\}_{i=1}^N$); *and* (4) *for all* $t\in\mathbb{R}$ *the function* $m(x,t;\lambda):\mathbb{C}\setminus(\mathcal{Z}_d\cup\widehat{\Gamma})\to\mathrm{SL}(2,\mathbb{C})$ *solves the following RH problem:*

a. $m(x,t;\lambda)$ *is meromorphic for all* $\lambda\in\mathbb{C}\setminus\widehat{\Gamma}$;

b. $m(x,t;\lambda)$ *satisfies the following jump conditions:*

$$m_+(x,t;\lambda)=m_-(x,t;\lambda)v(x,t;\lambda),\qquad\lambda\in\widehat{\Gamma},$$

where

$$v(x,t;\lambda):=\exp\{-i(\lambda^2x+2\lambda^4t)\mathrm{ad}(\sigma_3)\}\begin{pmatrix}1-r(\lambda)\overline{r(\overline{\lambda})} & r(\lambda)\\ -\overline{r(\overline{\lambda})} & 1\end{pmatrix},$$

$r(\lambda)$, *the reflection coefficient associated with the direct scattering problem for the operator* $\partial_x-U(x,t;\lambda)$, *satisfies* $r(-\lambda)=-r(\lambda)$, *and* $r(\lambda)\in\mathcal{S}(\widehat{\Gamma};\mathbb{C})$;

c. *for the simple poles of* $m(x,t;\lambda)$ *at* $\{\pm\lambda_j\}_{j=1}^N$ *and* $\{\pm\overline{\lambda_j}\}_{j=1}^N$, *there exist nilpotent matrices* $\{v_j(x,t)\sigma_-\}_{j=1}^N$ *and* $\{\overline{v_j(x,t)}\sigma_+\}_{j=1}^N$, *resp., such that the residues, for* $1\le j\le N$, *satisfy the* (*Beals–Coifman* [24, 25]) *polar conditions*

$$\begin{aligned}
\operatorname{res}(m(x,t;\lambda);\lambda_j) &= \lim_{\lambda\to\lambda_j} m(x,t;\lambda)v_j(x,t)\sigma_-,\\
\operatorname{res}(m(x,t;\lambda);-\lambda_j) &= -\sigma_3\operatorname{res}(m(x,t;\lambda);\lambda_j)\sigma_3,\\
\operatorname{res}(m(x,t;\lambda);\overline{\lambda_j}) &= \lim_{\lambda\to\overline{\lambda_j}} m(x,t;\lambda)\overline{v_j(x,t)}\sigma_+,\\
\operatorname{res}(m(x,t;\lambda);-\overline{\lambda_j}) &= -\sigma_3\operatorname{res}(m(x,t;\lambda);\overline{\lambda_j})\sigma_3,
\end{aligned}$$

where $v_j(x,t):=C_j\exp\{2i(\lambda_j^2x+2\lambda_j^4t)\}$, *and* C_j *are complex constants associated with the direct scattering problem for the operator* $\partial_x-U(x,t;\lambda)$;

d. *as* $\lambda\to\infty$, $\lambda\in\mathbb{C}\setminus(\mathcal{Z}_d\cup\widehat{\Gamma})$,

$$m(x,t;\lambda)=\mathrm{I}+\mathcal{O}(\lambda^{-1}).$$

Proof. Conditions (1)–(4) follow from the results given in [15, 24, 29, 30]. □

LEMMA 2.2. *Let* $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})}:=\sup_{\lambda\in\mathbb{R}}|r(\lambda)|<1$. *Then* (1) *the RH problem formulated in Lemma* 2.1 *is uniquely solvable*; (2) $\Psi(x,t;\lambda)=m(x,t;\lambda)\exp\{-i(\lambda^2x+2\lambda^4t)\sigma_3\}$ *is the solution of system* (5) *with*

$$Q(x,t):=2i\lim_{\lambda\to\infty}(\lambda m(x,t;\lambda))_{12}; \tag{11}$$

(3) *the function* $Q(x,t)$ *defined by* (11) *satisfies equation* (3), *and*

$$q(x,t):=Q(x,t)((m^{-1}(x,t;0))_{11})^2 \tag{12}$$

satisfies the DNLSE; *and* (4) $m(x,t;\lambda)$ *possesses the following symmetry reductions:* $m(x,t;\lambda)=\sigma_3m(x,t;-\lambda)\sigma_3$ *and* $m(x,t;\lambda)=\sigma_1\overline{m(x,t;\overline{\lambda})}\sigma_1$.

If $r(\lambda)\in\mathcal{S}(\widehat{\Gamma};\mathbb{C})$, *then, for any* $t\in\mathbb{R}$, $Q(x,t)$ (*resp.*, $q(x,t)$), *as a function of* x, $\in\mathcal{S}(\mathbb{R};\mathbb{C})$.

Proof. The solvability of the RH problem (formulated in Lemma 2.1) is a consequence of Theorem 9.3 in [31] and the vanishing winding number of $1-r(\lambda)\overline{r(\overline{\lambda})}$, $\int_{\widehat{\Gamma}}d(\arg(1-r(\lambda)\overline{r(\overline{\lambda})}))=\sum_{l\in\{\substack{>\\<}\}}\mathrm{s}(l)\mathrm{n}(l)=0$, where $\mathrm{s}(>)=-\mathrm{s}(<)=1$, and $\mathrm{n}(\substack{>\\<}):=\operatorname{card}(\{\lambda_j;\ \Im(\lambda_j^2)\substack{>\\<}0\})$; items (2) and (4) can be verified through straightforward calculations, and the fact that $q(x,t)$ (equation (12)) satisfies the DNLSE follows from Proposition 2.3 and the definition of $m(x,t;\lambda)$. □

Remark 2.2. In fact, in this paper, the solvability of the RH problem for $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})}<1$ is proved for all sufficiently large $|t|$: the solvability of the RH problem for $||r||_{\mathcal{L}^\infty(\widehat{\Gamma})}<1$ in the solitonless sector, $\mathcal{Z}_d\equiv\emptyset$, for all sufficiently large $|t|$ was proved in [15]. *Note:* the condition $||r||_{\mathcal{L}^\infty(\widehat{\Gamma})}<1$, which appears in [15], is restrictive and can be replaced by the weaker condition $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})}<1$.

Before summarizing the main results of this paper, namely, Theorems 2.1–2.3, some further preamble is required: (1) *the Kaup–Newell parametrization* [28] *is adopted for the discrete eigenvalues,* $\lambda_j:=\Delta_j\exp\{\frac{i}{2}(\pi-\gamma_j)\}$, $\Delta_j>0$, $\gamma_j\in(0,\pi)$, $1\le j\le N$, *and* $\lambda_j^2:=\xi_j+i\eta_j$, *where* $\xi_j=-\Delta_j^2\cos\gamma_j$ *and* $\eta_j=\Delta_j^2\sin\gamma_j$ (*note that, with this parametrization,* $\{\pm\lambda_i\}_{i=1}^N$ (*resp.*, $\{\pm\overline{\lambda_i}\}_{i=1}^N$) *lie in the* 1*st and* 3*rd quadrants* (*resp.*, 2*nd and* 4*th quadrants*) *of the complex plane of the auxiliary spectral parameter,* λ); *and* (2) *it is*

supposed throughout that $\xi_i \neq \xi_j$, $1 \leq i \neq j \leq N$, *so that it is convenient to choose the following enumeration for the points of the discrete spectrum (ordering of the solitons),* $\xi_1 > \cdots > \xi_n > \cdots > \xi_N$.

Remark 2.3. Even though the "symbol" ("notation") $C(z)$ appearing in the various *final* error estimations is *not* the same and should be properly denoted as $C_1(z)$, $C_2(z)$, etc., the simplified "notation" $C(z)$ is retained throughout since the principal concern here is not its explicit functional z-dependence but rather the functional class(es) to which it belongs. Throughout the paper, $M \in \mathbb{R}_{>0}$ is a fixed constant.

Remark 2.4. In Theorems 2.1–2.3 (see below), one should keep the upper signs as $t \to +\infty$ and the lower signs as $t \to -\infty$ everywhere.

THEOREM 2.1. *Let* $m(x,t;\lambda)$ *be the solution of the RH problem formulated in Lemma* 2.1 *with the condition* $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})} < 1$ *and* $Q(x,t)$ *be defined by* (11). *Then as* $t \to \pm\infty$ *and* $x \to \mp\infty$ *such that* $\lambda_0 := \frac{1}{2}\sqrt{-\frac{x}{t}} > M$ *and* $(x,t) \in \Omega_n := \{(x,t);\, x - 4t\Delta_n^2 \cos\gamma_n := l_n(t) = \mathcal{O}(1)\}$, *for those* $\gamma_n \in (\frac{\pi}{2}, \pi)$,

$$Q(x,t) = Q^{\pm}_{\mathrm{as}}(x,t) + \mathcal{O}\left(\frac{C(\lambda_0)\ln|t|}{t}\right), \tag{13}$$

where

$$Q^{\pm}_{\mathrm{as}}(x,t) := Q^{\mathcal{S}}_{\pm}(x,t) + Q^{\mathcal{C}}_{\pm}(x,t) + Q^{\mathcal{SC}}_{\pm}(x,t), \tag{14}$$

with

$$Q^{\mathcal{S}}_{\pm}(x,t) = \frac{2i\Delta_n \sin(\gamma_n)\exp\{\frac{i\gamma_n}{2}\}\exp\{2i(\Delta_n^2(2t\Delta_n^2 + l_n(t)\cos\gamma_n) + \widehat{\phi}^{\pm}_n)\}}{\cosh(\frac{i\gamma_n}{2} + 2\Delta_n^2 \sin(\gamma_n) l_n(t) - \widehat{x}^{\pm}_n)}, \tag{15}$$

$$\widehat{\phi}^{\pm}_n = -\frac{1}{2}\arg C_n + \arg\delta^{\pm}(\lambda_n;\lambda_0) + \sum_{l \in L_{\pm}} \arg\left(\frac{(\lambda_n - \overline{\lambda_l})(\lambda_n + \overline{\lambda_l})}{(\lambda_n - \lambda_l)(\lambda_n + \lambda_l)}\right), \tag{16}$$

$$\begin{aligned} \widehat{x}^{\pm}_n = &-\ln(\Delta_n \sin\gamma_n) + \ln|C_n| - 2\ln|\delta^{\pm}(\lambda_n;\lambda_0)| \\ &+ 2\sum_{l \in L_{\pm}} \ln\left(\frac{|\lambda_n - \lambda_l||\lambda_n + \lambda_l|}{|\lambda_n - \overline{\lambda_l}||\lambda_n + \overline{\lambda_l}|}\right), \end{aligned} \tag{17}$$

$$\delta^{+}(\lambda;z) = \exp\left\{\int_0^z \frac{\varrho \ln(1 - |r(\varrho)|^2)}{(\varrho^2 - \lambda^2)}\frac{d\varrho}{\pi i} - \int_0^{\infty} \frac{\varrho \ln(1 + |r(i\varrho)|^2)}{(\varrho^2 + \lambda^2)}\frac{d\varrho}{\pi i}\right\}, \tag{18}$$

$$\delta^{-}(\lambda;z) = \exp\left\{\int_z^{\infty} \frac{\varrho \ln(1 - |r(\varrho)|^2)}{(\varrho^2 - \lambda^2)}\frac{d\varrho}{\pi i}\right\}, \tag{19}$$

$$Q^{\mathcal{C}}_{\pm}(x,t) = \sqrt{\pm\frac{\nu(\lambda_0)}{2\lambda_0^2 t}}\exp\left\{i\left(\phi^{\pm}(\lambda_0) + \widehat{\Phi}^{\pm}(\lambda_0;t) + \frac{\pi}{2}\right)\right\}, \tag{20}$$

$$\nu(z) = -\frac{1}{2\pi}\ln(1 - |r(z)|^2), \tag{21}$$

$$\phi^+(z)=\frac{1}{\pi}\int_0^z \ln|\varrho^2-z^2|d\ln(1-|r(\varrho)|^2)$$
$$-\frac{1}{\pi}\int_0^\infty \ln|\varrho^2+z^2|d\ln(1+|r(i\varrho)|^2), \tag{22}$$

$$\phi^-(z)=\frac{1}{\pi}\int_z^\infty \ln|\varrho^2-z^2|d\ln(1-|r(\varrho)|^2), \tag{23}$$

$$\widehat{\Phi}^\pm(\lambda_0;t)=4\lambda_0^4 t\mp\nu(\lambda_0)\ln|t|\pm\arg\Gamma(i\nu(\lambda_0))+\arg r(\lambda_0)\mp 3\nu(\lambda_0)\ln 2$$
$$+\ (2\pm 1)\frac{\pi}{4}+2\sum_{l\in L_\pm}\arg\biggl(\frac{(\lambda_0-\overline{\lambda_l})(\lambda_0+\overline{\lambda_l})}{(\lambda_0-\lambda_l)(\lambda_0+\lambda_l)}\biggr), \tag{24}$$

$$Q_\pm^{\mathcal{SC}}(x,t)=-\frac{4(\Xi^\pm)^2\overline{g_n^\pm}|g_n^\pm|}{\eta_n}\sqrt{\pm\frac{\nu(\lambda_0)}{2\lambda_0^2 t}}\{\exp(i\varphi_n^\pm(\lambda_0;t))$$
$$+2i\cot(\gamma_n)\cos(\varphi_n^\pm(\lambda_0;t))\}, \tag{25}$$

$$g_n^\pm:=|g_n^\pm|\exp\{i\arg g_n^\pm\},$$

$$|g_n^\pm|=|C_n||\delta^\pm(\lambda_n;\lambda_0)|^{-2}\exp\{-2\Delta_n^2\sin(\gamma_n)l_n(t)\}$$
$$\cdot\ \exp\biggl\{2\sum_{l\in L_\pm}\ln\biggl(\frac{|\lambda_n-\lambda_l||\lambda_n+\lambda_l|}{|\lambda_n-\overline{\lambda_l}||\lambda_n+\overline{\lambda_l}|}\biggr)\biggr\}, \tag{26}$$

$$\arg g_n^\pm=\arg C_n-2\arg\delta^\pm(\lambda_n;\lambda_0)+2\sum_{l\in L_\pm}\arg\biggl(\frac{(\lambda_n-\lambda_l)(\lambda_n+\lambda_l)}{(\lambda_n-\overline{\lambda_l})(\lambda_n+\overline{\lambda_l})}\biggr)$$
$$-2\Delta_n^2(2t\Delta_n^2+l_n(t)\cos\gamma_n), \tag{27}$$

$$\Xi^\pm=\frac{\exp\{\frac{i\gamma_n}{2}\}\exp\{2\Delta_n^2\sin(\gamma_n)l_n(t)-\widehat{x}_n^\pm\}}{2\cosh(\frac{i\gamma_n}{2}+2\Delta_n^2\sin(\gamma_n)l_n(t)-\widehat{x}_n^\pm)}, \tag{28}$$

$$\varphi_n^\pm(\lambda_0;t):=\arg g_n^\pm+\phi^\pm(\lambda_0)+\widehat{\Phi}^\pm(\lambda_0;t), \tag{29}$$

$\sum_{l\in L_+}:=\sum_{l=n+1}^N$, $\sum_{l\in L_-}:=\sum_{l=1}^{n-1}$, $\Gamma(\cdot)$ *is the gamma function* [32], *and* $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$, *and, as* $t\to\pm\infty$ *and* $x\to\pm\infty$ *such that* $\mu_0:=\frac{1}{2}\sqrt{\frac{x}{t}}>M$ *and* $(x,t)\in\mathcal{O}_n:=\{(x,t);\ -x+4t\Delta_n^2\cos\gamma_n:=-l_n(t)=\mathcal{O}(1)\}$, *for those* $\gamma_n\in(0,\frac{\pi}{2})$,

$$Q(x,t)=Q_{\rm as}^{\pm\prime}(x,t)+\mathcal{O}\biggl(\frac{C(\mu_0)\ln|t|}{t}\biggr), \tag{30}$$

where

$$Q_{\rm as}^{\pm\prime}(x,t):=Q_\pm^{\mathcal{S}\prime}(x,t)+Q_\pm^{\mathcal{C}\prime}(x,t)+Q_\pm^{\mathcal{SC}\prime}(x,t), \tag{31}$$

with

$$Q_\pm^{\mathcal{S}\prime}(x,t)=\frac{2\Delta_n\sin(\gamma_n)\exp\{-\frac{i\gamma_n}{2}\}\exp\{2i(\Delta_n^2(2t\Delta_n^2+l_n(t)\cos\gamma_n)+\widehat{\phi}_n^{\pm\prime})\}}{\sinh(\frac{i\gamma_n}{2}+2\Delta_n^2\sin(\gamma_n)l_n(t)+\widehat{x}_n^{\pm\prime})}, \tag{32}$$

$$\widehat{\phi}_n^{\pm\prime} = -\frac{1}{2}\arg C_n + \arg\delta_\flat^{\pm}(\overline{\lambda_n};\mu_0) - \sum_{l\in L_\pm}\arg\left(\frac{(\lambda_n-\overline{\lambda_l})(\lambda_n+\overline{\lambda_l})}{(\lambda_n-\lambda_l)(\lambda_n+\lambda_l)}\right), \tag{33}$$

$$\begin{aligned}\widehat{x}_n^{\pm\prime} = &-\ln(\Delta_n\sin\gamma_n) + \ln|C_n| - 2\ln|\delta_\flat^{\pm}(\overline{\lambda_n};\mu_0)| \\ &+ 2\sum_{l\in L_\pm}\ln\left(\frac{|\lambda_n-\lambda_l||\lambda_n+\lambda_l|}{|\lambda_n-\overline{\lambda_l}||\lambda_n+\overline{\lambda_l}|}\right),\end{aligned} \tag{34}$$

$$\delta_\flat^{+}(\lambda;z) = \exp\left\{\int_0^z \frac{\varrho\ln(1+|r(i\varrho)|^2)}{(\varrho^2-\lambda^2)}\frac{d\varrho}{\pi i} - \int_0^\infty \frac{\varrho\ln(1-|r(\varrho)|^2)}{(\varrho^2+\lambda^2)}\frac{d\varrho}{\pi i}\right\}, \tag{35}$$

$$\delta_\flat^{-}(\lambda;z) = \exp\left\{\int_z^\infty \frac{\varrho\ln(1+|r(i\varrho)|^2)}{(\varrho^2-\lambda^2)}\frac{d\varrho}{\pi i}\right\}, \tag{36}$$

$$Q_\pm^{\mathcal{C}\prime}(x,t) = \sqrt{\mp\frac{\nu(i\mu_0)}{2\mu_0^2 t}}\exp\{i(\phi^{\pm\prime}(\mu_0)+\widehat{\Phi}^{\pm\prime}(\mu_0;t)+\pi)\}, \tag{37}$$

$$\nu(iz) = -\frac{1}{2\pi}\ln(1+|r(iz)|^2), \tag{38}$$

$$\begin{aligned}\phi^{+\prime}(z) = &\frac{1}{\pi}\int_0^z \ln|\varrho^2-z^2|d\ln(1+|r(i\varrho)|^2) \\ &-\frac{1}{\pi}\int_0^\infty \ln|\varrho^2+z^2|d\ln(1-|r(\varrho)|^2),\end{aligned} \tag{39}$$

$$\phi^{-\prime}(z) = \frac{1}{\pi}\int_z^\infty \ln|\varrho^2-z^2|d\ln(1+|r(i\varrho)|^2), \tag{40}$$

$$\begin{aligned}\widehat{\Phi}^{\pm\prime}(\mu_0;t) = &\,4\mu_0^4 t \mp \nu(i\mu_0)\ln|t| \pm \arg\Gamma(i\nu(i\mu_0)) + \arg r(i\mu_0) \mp 3\nu(i\mu_0)\ln 2 \\ &-\ (2\mp 1)\frac{\pi}{4} - 2\sum_{l\in L_\pm}\arg\left(\frac{(\mu_0-\overline{\lambda_l})(\mu_0+\overline{\lambda_l})}{(\mu_0-\lambda_l)(\mu_0+\lambda_l)}\right),\end{aligned} \tag{41}$$

$$\begin{aligned}Q_\pm^{\mathcal{SC}\prime}(x,t) = &-\frac{4i(\Xi^{\pm\prime})^2\overline{g_n^{\pm\prime}}|g_n^{\pm\prime}|}{\eta_n}\sqrt{\mp\frac{\nu(i\mu_0)}{2\mu_0^2 t}}\{\exp(i\varphi_n^{\pm\prime}(\mu_0;t)) \\ &-2i\cot(\gamma_n)\cos(\varphi_n^{\pm\prime}(\mu_0;t))\},\end{aligned} \tag{42}$$

$$g_n^{\pm\prime} := |g_n^{\pm\prime}|\exp\{i\arg g_n^{\pm\prime}\},$$

$$\begin{aligned}|g_n^{\pm\prime}| = &\,|C_n||\delta_\flat^{\pm}(\overline{\lambda_n};\mu_0)|^{-2}\exp\{2\Delta_n^2\sin(\gamma_n)l_n(t)\} \\ &\cdot\ \exp\left\{2\sum_{l\in L_\pm}\ln\left(\frac{|\lambda_n-\lambda_l||\lambda_n+\lambda_l|}{|\lambda_n-\overline{\lambda_l}||\lambda_n+\overline{\lambda_l}|}\right)\right\},\end{aligned} \tag{43}$$

$$\arg g_n^{\pm\prime} = \arg C_n - 2\arg\delta_\flat^\pm(\overline{\lambda_n};\mu_0) - 2\sum_{l\in L_\pm}\arg\left(\frac{(\lambda_n-\lambda_l)(\lambda_n+\lambda_l)}{(\lambda_n-\overline{\lambda_l})(\lambda_n+\overline{\lambda_l})}\right)$$

$$-\ 2\Delta_n^2(2t\Delta_n^2 + l_n(t)\cos\gamma_n), \tag{44}$$

$$\Xi^{\pm\prime} = -\frac{\exp\{-\frac{i\gamma_n}{2}\}\exp\{-2\Delta_n^2\sin(\gamma_n)l_n(t) - \widehat{x}_n^{\pm\prime}\}}{2\sinh(\frac{i\gamma_n}{2} + 2\Delta_n^2\sin(\gamma_n)l_n(t) + \widehat{x}_n^{\pm\prime})}, \tag{45}$$

$$\varphi_n^{\pm\prime}(\mu_0;t) := \arg g_n^{\pm\prime} + \phi^{\pm\prime}(\mu_0) + \widehat{\Phi}^{\pm\prime}(\mu_0;t), \tag{46}$$

and $C(\mu_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

THEOREM 2.2. *Let* $m(x,t;\lambda)$ *be the solution of the RH problem formulated in Lemma* 2.1 *with the condition* $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})}<1$ *and* $q(x,t)$, *the solution of the DNLSE (equation* (4)*) be defined by* (12) *in terms of the function* $Q(x,t)$ *given in Theorem* 2.1. *Then as* $t\to\pm\infty$ *and* $x\to\mp\infty$ *such that* $\lambda_0 := \frac{1}{2}\sqrt{-\frac{x}{t}} > M$ *and* $(x,t)\in\Omega_n :=$ $\{(x,t);\ x-4t\Delta_n^2\cos\gamma_n := l_n(t) = \mathcal{O}(1)\}$, *for those* $\gamma_n\in(\frac{\pi}{2},\pi)$,

$$q(x,t) = Q_{\rm as}^\pm(x,t)\exp\{i\arg q_{\rm as}^\pm(x,t)\} + \mathcal{O}\left(\frac{C(\lambda_0)(\ln|t|)^2}{t}\right), \tag{47}$$

where $Q_{\rm as}^\pm(x,t)$ *are given in Theorem* 2.1, (14)–(29),

$$\arg q_{\rm as}^\pm(x,t) = -4\sum_{l\in L_\pm}\gamma_l + 4\arctan(\eta_n|g_n^\pm|^{-2} + \cot\gamma_n) + \mathcal{Y}_\pm(\lambda_0)$$

$$+\ 4\sqrt{\pm\frac{\nu(\lambda_0)}{2\lambda_0^2 t}}\,\frac{|g_n^\pm|\sin(\gamma_n)(|g_n^\pm|^2\cos(\varphi_n^\pm(\lambda_0;t)-\gamma_n) - \eta_n\sin(\gamma_n)\cos(\varphi_n^\pm(\lambda_0;t)))}{((\eta_n\sin\gamma_n + |g_n^\pm|^2\cos\gamma_n)^2 + |g_n^\pm|^4\sin^2\gamma_n)}$$

$$-\sqrt{\pm\frac{2}{t}}\int_{\lambda_0}^\infty\frac{\sqrt{\nu(\mu)}}{\mu^2}\Big(\Re\{R^\pm(0)\}\cos(\widehat{\Theta}^\pm(\mu;t)) + \Im\{R^\pm(0)\}\sin(\widehat{\Theta}^\pm(\mu;t))\Big)\frac{d\mu}{\pi}, \tag{48}$$

$$\mathcal{Y}_+(z) = \frac{2}{\pi}\int_0^z\frac{\ln(1-|r(\varrho)|^2)}{\varrho}d\varrho - \frac{2}{\pi}\int_0^\infty\frac{\ln(1+|r(i\varrho)|^2)}{\varrho}d\varrho,$$

$$\mathcal{Y}_-(z) = \frac{2}{\pi}\int_z^\infty\frac{\ln(1-|r(\varrho)|^2)}{\varrho}d\varrho, \tag{49}$$

$$R^\pm(0) := \left(\left.\frac{d(r(z)|_{z\in\mathbb{R}})}{dz}\right|_{z=0} - \left.\frac{d(r(z)|_{z\in i\mathbb{R}})}{dz}\right|_{z=0}\right)\exp\left\{4i\sum_{l\in L_\pm}\gamma_l\right\}, \tag{50}$$

$$\widehat{\Theta}^\pm(\lambda_0;t) := \widehat{\Phi}^\pm(\lambda_0;t) + \phi^\pm(\lambda_0) + \mathcal{Y}_\pm(\lambda_0), \tag{51}$$

with $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$, *and, as* $t\to\pm\infty$ *and* $x\to\pm\infty$ *such that* $\mu_0 := \frac{1}{2}\sqrt{\frac{x}{t}} > M$ *and* $(x,t)\in\mathcal{U}_n := \{(x,t);\ -x+4t\Delta_n^2\cos\gamma_n := -l_n(t) = \mathcal{O}(1)\}$, *for those* $\gamma_n\in(0,\frac{\pi}{2})$,

$$q(x,t) = Q_{\rm as}^{\pm\prime}(x,t)\exp\{i\arg q_{\rm as}^{\pm\prime}(x,t)\} + \mathcal{O}\left(\frac{C(\mu_0)(\ln|t|)^2}{t}\right), \tag{52}$$

where $Q_{\rm as}^{\pm\prime}(x,t)$ *are given in Theorem* 2.1, (31)–(46),

$$\arg q_{\rm as}^{\pm\prime}(x,t) = 4\sum_{l\in L_\pm}\gamma_l + 4\arctan(\eta_n|g_n^{\pm\prime}|^{-2}-\cot\gamma_n) + \mathcal{Y}'_\pm(\mu_0)$$
$$- 4\sqrt{\mp\frac{\nu(i\mu_0)}{2\mu_0^2 t}}\,\frac{|g_n^{\pm\prime}|\sin(\gamma_n)(|g_n^{\pm\prime}|^2\cos(\varphi_n^{\pm\prime}(\mu_0;t)+\gamma_n)+\eta_n\sin(\gamma_n)\cos(\varphi_n^{\pm\prime}(\mu_0;t)))}{((\eta_n\sin\gamma_n-|g_n^{\pm\prime}|^2\cos\gamma_n)^2+|g_n^{\pm\prime}|^4\sin^2\gamma_n)}$$
$$- \sqrt{\pm\frac{2}{t}}\int_{\mu_0}^{\infty}\frac{\sqrt{-\nu(i\mu)}}{\mu^2}$$

$$\cdot\Big(\Re\{R^{\pm\prime}(0)\}\cos(\widehat{\Theta}^{\pm\prime}(\mu;t))+\Im\{R^{\pm\prime}(0)\}\sin(\widehat{\Theta}^{\pm\prime}(\mu;t))\Big)\frac{d\mu}{\pi}, \tag{53}$$

$$\mathcal{Y}'_+(z) = \frac{2}{\pi}\int_0^z\frac{\ln(1+|r(i\varrho)|^2)}{\varrho}d\varrho - \frac{2}{\pi}\int_0^\infty\frac{\ln(1-|r(\varrho)|^2)}{\varrho}d\varrho,$$

$$\mathcal{Y}'_-(z) = \frac{2}{\pi}\int_z^\infty\frac{\ln(1+|r(i\varrho)|^2)}{\varrho}d\varrho, \tag{54}$$

$$R^{\pm\prime}(0) := \Big(\left.\frac{d(r(z)|_{z\in\mathbb{R}})}{dz}\right|_{z=0} - \left.\frac{d(r(z)|_{z\in i\mathbb{R}})}{dz}\right|_{z=0}\Big)\exp\Big\{-4i\sum_{l\in L_\pm}\gamma_l\Big\}, \tag{55}$$

$$\widehat{\Theta}^{\pm\prime}(\mu_0;t) := \widehat{\Phi}^{\pm\prime}(\mu_0;t) + \phi^{\pm\prime}(\mu_0) + \mathcal{Y}'_\pm(\mu_0), \tag{56}$$

and $C(\mu_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

THEOREM 2.3. *Let* $m(x,t;\lambda)$ *be the solution of the RH problem formulated in Lemma* 2.1 *with the condition* $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})}<1$ *and* $u(x,t)$, *the solution of the MNLSE (equation* (2)*), be defined by* (10) *in terms of the function* $q(x,t)$ *given in Theorem* 2.2. *Then as* $t\to\pm\infty$ *and* $x\to\pm\infty$ *such that* $\widehat{\lambda}_0 := \sqrt{\frac{1}{2}(\frac{x}{t}-\frac{1}{s})} > M$, $\frac{x}{t}>\frac{1}{s}$, $s\in\mathbb{R}_{>0}$, *and* $(x,t)\in\widetilde{\Omega}_n := \{(x,t);\ -x+t(\frac{1}{s}-2\Delta_n^2\cos\gamma_n) := \widehat{l}_n(t) = \mathcal{O}(1)\}$, *for those* $\gamma_n\in(\frac{\pi}{2},\pi)$,

$$u(x,t) = v_{\rm as}^{\pm}(x,t)w_{\rm as}^{\pm}(x,t) + \mathcal{O}\left(\frac{C(\widehat{\lambda}_0)(\ln|t|)^2}{t}\right), \tag{57}$$

where

$$v_{\rm as}^{\pm}(x,t) := v_\pm^{\mathcal{S}}(x,t) + v_\pm^{\mathcal{C}}(x,t) + v_\pm^{\mathcal{SC}}(x,t), \tag{58}$$

with

$$v_\pm^{\mathcal{S}}(x,t) = \frac{\sqrt{2}i\Delta_n\sin(\gamma_n)\exp\{\frac{i\gamma_n}{2}\}\exp\{2i(\Delta_n^2(t\Delta_n^2+\widehat{l}_n(t)\cos\gamma_n)+\widetilde{\phi}_n^{\pm})\}}{\sqrt{s}\cosh(\frac{i\gamma_n}{2}+2\Delta_n^2\sin(\gamma_n)\widehat{l}_n(t)-\widetilde{x}_n^{\pm})}, \tag{59}$$

$$\widetilde{\phi}_n^{\pm} = -\frac{1}{2}\arg C_n + \arg\delta^{\pm}(\lambda_n;\widehat{\lambda}_0) + \sum_{l\in L_\pm}\arg\left(\frac{(\lambda_n-\overline{\lambda_l})(\lambda_n+\overline{\lambda_l})}{(\lambda_n-\lambda_l)(\lambda_n+\lambda_l)}\right), \tag{60}$$

$$\widetilde{x}_n^{\pm} = -\ln(\Delta_n \sin\gamma_n) + \ln|C_n| - 2\ln|\delta^{\pm}(\lambda_n;\widehat{\lambda}_0)| + 2\sum_{l\in L_{\pm}} \ln\left(\frac{|\lambda_n-\lambda_l||\lambda_n+\lambda_l|}{|\lambda_n-\overline{\lambda_l}||\lambda_n+\overline{\lambda_l}|}\right), \tag{61}$$

$$v_{\pm}^{\mathcal{C}}(x,t) = \sqrt{\pm\frac{\nu(\widehat{\lambda}_0)}{2\widehat{\lambda}_0^2 st}}\exp\left\{i\left(\phi^{\pm}(\widehat{\lambda}_0)+\widetilde{\Phi}^{\pm}(\widehat{\lambda}_0;t)+\frac{\pi}{2}\right)\right\}, \tag{62}$$

$$\begin{aligned}\widetilde{\Phi}^{\pm}(\widehat{\lambda}_0;t) &= 2\widehat{\lambda}_0^4 t \mp \nu(\widehat{\lambda}_0)\ln|t| \pm \arg\Gamma(i\nu(\widehat{\lambda}_0)) + \arg r(\widehat{\lambda}_0) \mp 2\nu(\widehat{\lambda}_0)\ln 2 \\ &+ (2\pm 1)\frac{\pi}{4} + 2\sum_{l\in L_{\pm}} \arg\left(\frac{(\widehat{\lambda}_0-\overline{\lambda_l})(\widehat{\lambda}_0+\overline{\lambda_l})}{(\widehat{\lambda}_0-\lambda_l)(\widehat{\lambda}_0+\lambda_l)}\right),\end{aligned} \tag{63}$$

$$\begin{aligned}v_{\pm}^{\mathcal{SC}}(x,t) = -\frac{4(\widetilde{\Xi}^{\pm})^2\overline{\widetilde{g}_n^{\pm}}|\widetilde{g}_n^{\pm}|}{\eta_n}\sqrt{\pm\frac{\nu(\widehat{\lambda}_0)}{2\widehat{\lambda}_0^2 st}}\{&\exp(i\widetilde{\varphi}_n^{\pm}(\widehat{\lambda}_0;t)) \\ &+ 2i\cot(\gamma_n)\cos(\widetilde{\varphi}_n^{\pm}(\widehat{\lambda}_0;t))\},\end{aligned} \tag{64}$$

$$\begin{aligned}\widetilde{g}_n^{\pm} &:= |\widetilde{g}_n^{\pm}|\exp\{i\arg\widetilde{g}_n^{\pm}\}, \\ |\widetilde{g}_n^{\pm}| &= |C_n||\delta^{\pm}(\lambda_n;\widehat{\lambda}_0)|^{-2}\exp\{-2\Delta_n^2\sin(\gamma_n)\widehat{l}_n(t)\} \\ &\cdot\exp\left\{2\sum_{l\in L_{\pm}}\ln\left(\frac{|\lambda_n-\lambda_l||\lambda_n+\lambda_l|}{|\lambda_n-\overline{\lambda_l}||\lambda_n+\overline{\lambda_l}|}\right)\right\},\end{aligned} \tag{65}$$

$$\begin{aligned}\arg\widetilde{g}_n^{\pm} &= \arg C_n - 2\arg\delta^{\pm}(\lambda_n;\widehat{\lambda}_0) + 2\sum_{l\in L_{\pm}}\arg\left(\frac{(\lambda_n-\lambda_l)(\lambda_n+\lambda_l)}{(\lambda_n-\overline{\lambda_l})(\lambda_n+\overline{\lambda_l})}\right) \\ &- 2\Delta_n^2(t\Delta_n^2+\widehat{l}_n(t)\cos\gamma_n),\end{aligned} \tag{66}$$

$$\widetilde{\Xi}^{\pm} = \frac{\exp\{\frac{i\gamma_n}{2}\}\exp\{2\Delta_n^2\sin(\gamma_n)\widehat{l}_n(t)-\widetilde{x}_n^{\pm}\}}{2\cosh(\frac{i\gamma_n}{2}+2\Delta_n^2\sin(\gamma_n)\widehat{l}_n(t)-\widetilde{x}_n^{\pm})}, \tag{67}$$

$$\widetilde{\varphi}_n^{\pm}(\widehat{\lambda}_0;t) := \arg\widetilde{g}_n^{\pm} + \phi^{\pm}(\widehat{\lambda}_0) + \widetilde{\Phi}^{\pm}(\widehat{\lambda}_0;t), \tag{68}$$

$$\begin{aligned}&w_{\mathrm{as}}^{\pm}(x,t) \\ &:= \exp\Bigg\{i\Bigg(-4\sum_{l\in L_{\pm}}\gamma_l + 4\arctan(\eta_n|\widetilde{g}_n^{\pm}|^{-2}+\cot\gamma_n) + \mathcal{Y}_{\pm}(\widehat{\lambda}_0) + \frac{t}{2s^2}(4\widehat{\lambda}_0^2 s+1) \\ &\quad + 4\sqrt{\pm\frac{\nu(\widehat{\lambda}_0)}{\widehat{\lambda}_0^2 t}}\frac{|\widetilde{g}_n^{\pm}|\sin(\gamma_n)(|\widetilde{g}_n^{\pm}|^2\cos(\widetilde{\varphi}_n^{\pm}(\widehat{\lambda}_0;t)-\gamma_n)-\eta_n\sin(\gamma_n)\cos(\widetilde{\varphi}_n^{\pm}(\widehat{\lambda}_0;t)))}{((\eta_n\sin\gamma_n+|\widetilde{g}_n^{\pm}|^2\cos\gamma_n)^2+|\widetilde{g}_n^{\pm}|^4\sin^2\gamma_n)} \\ &\quad - \frac{2}{\sqrt{\pm t}}\int_{\widehat{\lambda}_0}^{\infty}\frac{\sqrt{\nu(\mu)}}{\mu^2}\Big(\Re\{R^{\pm}(0)\}\cos(\widetilde{\Theta}^{\pm}(\mu;t))+\Im\{R^{\pm}(0)\}\sin(\widetilde{\Theta}^{\pm}(\mu;t))\Big)\frac{d\mu}{\pi}\Bigg)\Bigg\},\end{aligned} \tag{69}$$

$$\widetilde{\Theta}^{\pm}(\widehat{\lambda}_0;t) := \widetilde{\Phi}^{\pm}(\widehat{\lambda}_0;t) + \phi^{\pm}(\widehat{\lambda}_0) + \mathcal{Y}_{\pm}(\widehat{\lambda}_0), \tag{70}$$

and $C(\widehat{\lambda}_0) \in \mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$, *and, as* $t \to \pm\infty$ *and* $x \to \mp\infty$ *or* $\pm\infty$ *such that* $\widehat{\mu}_0 := \sqrt{\frac{1}{2}(\frac{1}{s}-\frac{x}{t})} > M$, $\frac{x}{t} < \frac{1}{s}$, $s \in \mathbb{R}_{>0}$, *and* $(x,t) \in \widetilde{\mho}_n := \{(x,t);\, x - t(\frac{1}{s} - 2\Delta_n^2 \cos\gamma_n) := -\widehat{l}_n(t) = \mathcal{O}(1)\}$, *for those* $\gamma_n \in (0, \frac{\pi}{2})$,

$$u(x,t) = v_{\mathrm{as}}^{\pm\prime}(x,t) w_{\mathrm{as}}^{\pm\prime}(x,t) + \mathcal{O}\bigg(\frac{C(\widehat{\mu}_0)(\ln|t|)^2}{t}\bigg), \tag{71}$$

where

$$v_{\mathrm{as}}^{\pm\prime}(x,t) := v_{\pm}^{\mathcal{S}\prime}(x,t) + v_{\pm}^{\mathcal{C}\prime}(x,t) + v_{\pm}^{\mathcal{SC}\prime}(x,t), \tag{72}$$

with

$$v_{\pm}^{\mathcal{S}\prime}(x,t) = \frac{\sqrt{2}\Delta_n \sin(\gamma_n) \exp\{-\frac{i\gamma_n}{2}\} \exp\{2i(\Delta_n^2(t\Delta_n^2 + \widehat{l}_n(t)\cos\gamma_n) + \widetilde{\phi}_n^{\pm\prime})\}}{\sqrt{s}\sinh(\frac{i\gamma_n}{2} + 2\Delta_n^2 \sin(\gamma_n)\widehat{l}_n(t) + \widetilde{x}_n^{\pm\prime})}, \tag{73}$$

$$\widetilde{\phi}_n^{\pm\prime} = -\frac{1}{2}\arg C_n + \arg\delta_{\flat}^{\pm}(\overline{\lambda_n};\widehat{\mu}_0) - \sum_{l\in L_{\pm}} \arg\bigg(\frac{(\lambda_n - \overline{\lambda_l})(\lambda_n + \overline{\lambda_l})}{(\lambda_n - \lambda_l)(\lambda_n + \lambda_l)}\bigg), \tag{74}$$

$$\begin{aligned}\widetilde{x}_n^{\pm\,\prime} = &-\ln(\Delta_n \sin\gamma_n) + \ln|C_n| - 2\ln|\delta_{\flat}^{\pm}(\overline{\lambda_n};\widehat{\mu}_0)| \\ &+ 2\sum_{l\in L_{\pm}} \ln\bigg(\frac{|\lambda_n - \lambda_l||\lambda_n + \lambda_l|}{|\lambda_n - \overline{\lambda_l}||\lambda_n + \overline{\lambda_l}|}\bigg),\end{aligned} \tag{75}$$

$$v_{\pm}^{\mathcal{C}\prime}(x,t) = \sqrt{\mp\frac{\nu(i\widehat{\mu}_0)}{2\widehat{\mu}_0^2 st}} \exp\{i(\phi^{\pm\prime}(\widehat{\mu}_0) + \widetilde{\Phi}^{\pm\prime}(\widehat{\mu}_0;t) + \pi)\}, \tag{76}$$

$$\begin{aligned}\widetilde{\Phi}^{\pm\prime}(\widehat{\mu}_0;t) = {}& 2\widehat{\mu}_0^4 t \mp \nu(i\widehat{\mu}_0)\ln|t| \pm \arg\Gamma(i\nu(i\widehat{\mu}_0)) + \arg r(i\widehat{\mu}_0) \mp 2\nu(i\widehat{\mu}_0)\ln 2 \\ &- (2\mp 1)\frac{\pi}{4} - 2\sum_{l\in L_{\pm}} \arg\bigg(\frac{(\widehat{\mu}_0 - \overline{\lambda_l})(\widehat{\mu}_0 + \overline{\lambda_l})}{(\widehat{\mu}_0 - \lambda_l)(\widehat{\mu}_0 + \lambda_l)}\bigg),\end{aligned} \tag{77}$$

$$\begin{aligned}v_{\pm}^{\mathcal{SC}\prime}(x,t) = {}& -\frac{4i(\widetilde{\Xi}^{\pm\prime})^2 \overline{\widetilde{g}_n^{\pm\prime}}|\widetilde{g}_n^{\pm\prime}|}{\eta_n} \sqrt{\mp\frac{\nu(i\widehat{\mu}_0)}{2\widehat{\mu}_0^2 st}} \{\exp(i\widetilde{\varphi}_n^{\pm\prime}(\widehat{\mu}_0;t)) \\ &- 2i\cot(\gamma_n)\cos(\widetilde{\varphi}_n^{\pm\prime}(\widehat{\mu}_0;t))\},\end{aligned} \tag{78}$$

$$\begin{aligned}\widetilde{g}_n^{\pm\prime} &:= |\widetilde{g}_n^{\pm\prime}| \exp\{i\arg\widetilde{g}_n^{\pm\prime}\}, \\ |\widetilde{g}_n^{\pm\prime}| &= |C_n||\delta_{\flat}^{\pm}(\overline{\lambda_n};\widehat{\mu}_0)|^{-2} \exp\{2\Delta_n^2 \sin(\gamma_n)\widehat{l}_n(t)\} \\ &\quad \cdot \exp\bigg\{2\sum_{l\in L_{\pm}} \ln\bigg(\frac{|\lambda_n - \lambda_l||\lambda_n + \lambda_l|}{|\lambda_n - \overline{\lambda_l}||\lambda_n + \overline{\lambda_l}|}\bigg)\bigg\},\end{aligned} \tag{79}$$

$$\begin{aligned}\arg\widetilde{g}_n^{\pm\prime} = {}& \arg C_n - 2\arg\delta_{\flat}^{\pm}(\overline{\lambda_n};\widehat{\mu}_0) - 2\sum_{l\in L_{\pm}} \arg\bigg(\frac{(\lambda_n - \lambda_l)(\lambda_n + \lambda_l)}{(\lambda_n - \overline{\lambda_l})(\lambda_n + \overline{\lambda_l})}\bigg) \\ &- 2\Delta_n^2(t\Delta_n^2 + \widehat{l}_n(t)\cos\gamma_n),\end{aligned} \tag{80}$$

$$\widetilde{\Xi}^{\pm\prime} = -\frac{\exp\{-\frac{i\gamma_n}{2}\} \exp\{-2\Delta_n^2 \sin(\gamma_n)\widehat{l}_n(t) - \widetilde{x}_n^{\pm\prime}\}}{2\sinh(\frac{i\gamma_n}{2} + 2\Delta_n^2 \sin(\gamma_n)\widehat{l}_n(t) + \widetilde{x}_n^{\pm\prime})}, \tag{81}$$

$$\widetilde{\varphi}_n^{\pm\prime}(\widehat{\mu}_0;t) := \arg \widetilde{g}_n^{\pm\prime} + \phi^{\pm\prime}(\widehat{\mu}_0) + \widetilde{\Phi}^{\pm\prime}(\widehat{\mu}_0;t), \tag{82}$$

$$\begin{aligned} &w_{\mathrm{as}}^{\pm\prime}(x,t) \\ &:= \exp\Bigg\{ i \, \Bigg(4 \sum_{l\in L_\pm} \gamma_l + 4 \arctan(\eta_n |\widetilde{g}_n^{\pm\prime}|^{-2} - \cot\gamma_n) + \mathcal{Y}_\pm'(\widehat{\mu}_0) + \frac{t}{2s^2}(-4\widehat{\mu}_0^2 s + 1) \\ &\quad -4\sqrt{\mp \frac{\nu(i\widehat{\mu}_0)}{\widehat{\mu}_0^2 t}} \, \frac{|\widetilde{g}_n^{\pm\prime}| \sin(\gamma_n)(|\widetilde{g}_n^{\pm\prime}|^2 \cos(\widetilde{\varphi}_n^{\pm\prime}(\widehat{\mu}_0;t) + \gamma_n) + \eta_n \sin(\gamma_n) \cos(\widetilde{\varphi}_n^{\pm\prime}(\widehat{\mu}_0;t)))}{((\eta_n \sin\gamma_n - |\widetilde{g}_n^{\pm\prime}|^2 \cos\gamma_n)^2 + |\widetilde{g}_n^{\pm\prime}|^4 \sin^2\gamma_n)} \\ &\quad - \frac{2}{\sqrt{\pm t}} \int_{\widehat{\mu}_0}^{\infty} \frac{\sqrt{-\nu(i\mu)}}{\mu^2} \Big(\Re\{R^{\pm\prime}(0)\} \cos(\widetilde{\Theta}^{\pm\prime}(\mu;t)) \\ &\qquad\qquad + \Im\{R^{\pm\prime}(0)\} \sin(\widetilde{\Theta}^{\pm\prime}(\mu;t)) \Big) \frac{d\mu}{\pi} \Bigg) \Bigg\}, \end{aligned} \tag{83}$$

$$\widetilde{\Theta}^{\pm\prime}(\widehat{\mu}_0;t) := \widetilde{\Phi}^{\pm\prime}(\widehat{\mu}_0;t) + \phi^{\pm\prime}(\widehat{\mu}_0) + \mathcal{Y}_\pm'(\widehat{\mu}_0), \tag{84}$$

and $C(\widehat{\mu}_0) \in \mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

One possible application of the asymptotic results obtained in Theorems 2.1–2.3 is associated with the so-called "soliton scattering," namely, the calculation of the position and phase shifts of the nth soliton ($1 \le n \le N$) for $Q(x,t)$, $q(x,t)$, and $u(x,t)$ in the presence of the continuous spectrum: other physical applications of these asymptotic results include, for example, the calculation of the temporal and spectral intensities for the solutions of the DNLSE and MNLSE.

Corollary 2.1.

(A) $Q(x,t)$:

$$\begin{aligned} \Delta x_n^{Q^{\mathcal{S}}} &:= (2\eta_n)^{-1}(\widehat{x}_n^+ - \widehat{x}_n^-) \\ &= \eta_n^{-1} \left\{ \sum_{\substack{l=1 \\ \neq n}}^{N} \operatorname{sgn}(l-n) \ln\left(\frac{|\lambda_n - \lambda_l||\lambda_n + \lambda_l|}{|\lambda_n - \overline{\lambda_l}||\lambda_n + \overline{\lambda_l}|} \right) - \ln\left(\frac{|\delta^+(\lambda_n;\lambda_0)|}{|\delta^-(\lambda_n;\lambda_0)|} \right) \right\}, \end{aligned}$$

$$\begin{aligned} \Delta\phi_n^{Q^{\mathcal{S}}} &:= 2(\widehat{\phi}_n^+ - \widehat{\phi}_n^-) \\ &= 2 \left\{ \sum_{\substack{l=1 \\ \neq n}}^{N} \operatorname{sgn}(l-n) \arg\left(\frac{(\lambda_n - \overline{\lambda_l})(\lambda_n + \overline{\lambda_l})}{(\lambda_n - \lambda_l)(\lambda_n + \lambda_l)} \right) + \arg\left(\frac{\delta^+(\lambda_n;\lambda_0)}{\delta^-(\lambda_n;\lambda_0)} \right) \right\}, \end{aligned}$$

$$\begin{aligned} \Delta x_n^{Q^{\mathcal{S}\prime}} &:= (2\eta_n)^{-1}(\widehat{x}_n^{+\prime} - \widehat{x}_n^{-\prime}) \\ &= \eta_n^{-1} \left\{ \sum_{\substack{l=1 \\ \neq n}}^{N} \operatorname{sgn}(l-n) \ln\left(\frac{|\lambda_n - \lambda_l||\lambda_n + \lambda_l|}{|\lambda_n - \overline{\lambda_l}||\lambda_n + \overline{\lambda_l}|} \right) - \ln\left(\frac{|\delta_\flat^+(\overline{\lambda_n};\mu_0)|}{|\delta_\flat^-(\overline{\lambda_n};\mu_0)|} \right) \right\}, \end{aligned}$$

$$\begin{aligned} \Delta\phi_n^{Q^{\mathcal{S}\prime}} &:= 2(\widehat{\phi}_n^{+\prime} - \widehat{\phi}_n^{-\prime}) \\ &= -2 \left\{ \sum_{\substack{l=1 \\ \neq n}}^{N} \operatorname{sgn}(l-n) \arg\left(\frac{(\lambda_n - \overline{\lambda_l})(\lambda_n + \overline{\lambda_l})}{(\lambda_n - \lambda_l)(\lambda_n + \lambda_l)} \right) - \arg\left(\frac{\delta_\flat^+(\overline{\lambda_n};\mu_0)}{\delta_\flat^-(\overline{\lambda_n};\mu_0)} \right) \right\}; \end{aligned}$$

(B) $q(x,t)$ (DNLSE):

$$\begin{aligned}
\Delta x_n^{q^{\mathcal{S}}} &= \Delta x_n^{Q^{\mathcal{S}}},\\
\Delta\phi_n^{q^{\mathcal{S}}} &= \Delta\phi_n^{Q^{\mathcal{S}}} - 4\sum_{\substack{l=1\\ \neq n}}^{N}\operatorname{sgn}(l-n)\gamma_l + \mathcal{Y}_+(\lambda_0) - \mathcal{Y}_-(\lambda_0),\\
\Delta x_n^{q^{\mathcal{S}\prime}} &= \Delta x_n^{Q^{\mathcal{S}\prime}},\\
\Delta\phi_n^{q^{\mathcal{S}\prime}} &= \Delta\phi_n^{Q^{\mathcal{S}\prime}} + 4\sum_{\substack{l=1\\ \neq n}}^{N}\operatorname{sgn}(l-n)\gamma_l + \mathcal{Y}'_+(\mu_0) - \mathcal{Y}'_-(\mu_0);
\end{aligned}$$

(C) $u(x,t)$ (MNLSE):

$$\begin{aligned}
\Delta x_n^{u^{\mathcal{S}}} &= \Delta x_n^{Q^{\mathcal{S}}}\big|_{\lambda_0\to\widehat{\lambda}_0},\\
\Delta\phi_n^{u^{\mathcal{S}}} &= \Delta\phi_n^{q^{\mathcal{S}}}\big|_{\lambda_0\to\widehat{\lambda}_0},\\
\Delta x_n^{u^{\mathcal{S}\prime}} &= \Delta x_n^{Q^{\mathcal{S}\prime}}\big|_{\mu_0\to\widehat{\mu}_0},\\
\Delta\phi_n^{u^{\mathcal{S}\prime}} &= \Delta\phi_n^{q^{\mathcal{S}\prime}}\big|_{\mu_0\to\widehat{\mu}_0}.
\end{aligned}$$

Proof. The proof follows from the definition of soliton position and phase shifts given in [4] and Theorems 2.1–2.3, equations (16), (17), (33), (34), (48), (53), (60), (61), (69), (74), (75), and (83). □

Remark 2.5. The expressions for the soliton phase shifts given in Corollary 2.1, namely, $\Delta\phi_n^{Q^{\mathcal{S}}}$, $\Delta\phi_n^{Q^{\mathcal{S}\prime}}$, $\Delta\phi_n^{q^{\mathcal{S}}}$, $\Delta\phi_n^{q^{\mathcal{S}\prime}}$, $\Delta\phi_n^{u^{\mathcal{S}}}$, and $\Delta\phi_n^{u^{\mathcal{S}\prime}}$, $1\le n\le N$, are to be understood $\operatorname{mod}(2\pi)$.

Remark 2.6. For the asymptotics of the $\mathbb{C}$-valued functions $Q(x,t)$, $q(x,t)$, and $u(x,t)$, one must actually consider four different cases, depending, resp., on the quadrant of the (x,t)-plane. In this paper, the proof of the asymptotic expansions for $Q(x,t)$ and $q(x,t)$ (resp., $u(x,t)$) is presented for the cases $(x,t)\to(\mp\infty,\pm\infty)$ (resp., $(x,t)\to(\pm\infty,\pm\infty)$) such that $\lambda_0 > M$ and $(x,t)\in\Omega_n$ (resp., $\widehat{\lambda}_0 > M$ and $(x,t)\in\widetilde{\Omega}_n$) for those $\gamma_n\in(\frac{\pi}{2},\pi)$: the results for the remaining domains of the (x,t)-plane are obtained analogously. If the conditions on γ_n stated in Theorems 2.1–2.3 are violated, then $(x,t)\in\{\mathbb{R}^2\setminus\Omega_n, \mathbb{R}^2\setminus\widetilde{\Omega}_n, \mathbb{R}^2\setminus\mho_n, \mathbb{R}^2\setminus\widetilde{\mho}_n\}$, but the asymptotic results stated still remain valid although the second terms on the right-hand sides of the asymptotic expansions become the leading-order terms of the corresponding asymptotic expansions, while the remaining terms are exponentially small and negligible with respect to the given error estimations.

3. The model RH problem. In order to simplify the asymptotic analysis of the original RH problem formulated in Lemma 2.1, a simpler, model RH problem (see Lemma 3.3) is derived in this section. As an intermediate step towards the formulation of the model RH problem, it will be convenient to derive an "extended" RH problem (see Lemma 3.2): the general idea pertaining to the transformations from the original RH problem to the model one is elucidated in the paragraph following Lemma 3.1 (see below).

PROPOSITION 3.1 (see [15]). *In the solitonless sector* ($\mathcal{Z}_d \equiv \emptyset$), *as* $t \to +\infty$ *and* $x \to -\infty$ *such that* $\lambda_0 := \frac{1}{2}\sqrt{-\frac{x}{t}} > M$,

$$m(x,t;\lambda) = \Delta(\lambda) + \mathcal{O}\left(\frac{C(\lambda_0)}{\sqrt{t}}\right),$$

where $\Delta(\lambda) := (\delta^+(\lambda;\lambda_0))^{\sigma_3}$,

$$\delta^+(\lambda;\lambda_0) = \left(\left(\frac{\lambda-\lambda_0}{\lambda}\right)\left(\frac{\lambda+\lambda_0}{\lambda}\right)\right)^{i\nu} \exp\left\{\sum_{l\in\{\pm\}} (\rho_l(\lambda) + \widehat{\rho}_l(\lambda))\right\},$$

$$\rho_\pm(\lambda) = \frac{1}{2\pi i}\int_0^{\pm\lambda_0} \ln\left(\frac{1-|r(\varsigma)|^2}{1-|r(\lambda_0)|^2}\right)\frac{d\varsigma}{(\varsigma-\lambda)}, \qquad \widehat{\rho}_\pm(\lambda) = \int_{\pm i\infty}^{i0} \frac{\ln(1-r(\varsigma)\overline{r(\bar{\varsigma})})}{(\varsigma-\lambda)}\frac{d\varsigma}{2\pi i},$$

$\nu := \nu(\lambda_0)$ *is given by* (21), $||(\delta^+(\cdot;\lambda_0))^{\pm 1}||_{\mathcal{L}^\infty(\mathbb{C};\mathbb{C})} := \sup_{\lambda\in\mathbb{C}} |(\delta^+(\lambda;\lambda_0))^{\pm 1}| < \infty$, $\overline{(\delta^+(\pm\overline{\lambda};\lambda_0))^{-1}} = \delta^+(\lambda;\lambda_0)$, *the principal branch of the logarithmic function is taken*, $\ln(\mu-\lambda) := \ln|\mu-\lambda| + i\arg(\mu-\lambda)$, $\arg(\mu-\lambda) \in (-\pi,\pi)$, *and* $C(\lambda_0) \in \mathcal{S}(\mathbb{R}_{>M}; M_2(\mathbb{C}))$.

Remark 3.1. For notational convenience, until the end of section 5, all explicit x,t dependencies are suppressed, except where absolutely necessary, and $\delta^+(\lambda;\lambda_0) := \delta(\lambda)$.

LEMMA 3.1. *There exists a unique solution* $m^\Delta(\lambda): \mathbb{C}\setminus(\mathcal{Z}_d \cup \widehat{\Gamma}) \to \mathrm{SL}(2,\mathbb{C})$ *of the following RH problem:*

1. $m^\Delta(\lambda)$ *is meromorphic for all* $\lambda \in \mathbb{C}\setminus\widehat{\Gamma}$,
2. $$m_+^\Delta(\lambda) = m_-^\Delta(\lambda) v^\Delta(\lambda), \quad \lambda \in \widehat{\Gamma},$$
   where
   $$v^\Delta(\lambda) = e^{-i\theta(\lambda)\mathrm{ad}(\sigma_3)} \begin{pmatrix} (1-r(\lambda)\overline{r(\overline{\lambda})})\delta_-(\lambda)(\delta_+(\lambda))^{-1} & r(\lambda)\delta_-(\lambda)\delta_+(\lambda) \\ -\overline{r(\overline{\lambda})}(\delta_-(\lambda))^{-1}(\delta_+(\lambda))^{-1} & (\delta_-(\lambda))^{-1}\delta_+(\lambda) \end{pmatrix},$$
   and $\theta(\lambda) := \lambda^2 x + 2\lambda^4 t$,
3. $m^\Delta(\lambda)$ *has simple poles at* $\{\pm\lambda_i, \pm\overline{\lambda_i}\}_{i=1}^N$ *with* $(1 \le i \le N)$,
   $$\begin{aligned} \mathrm{res}(m^\Delta(\lambda);\lambda_i) &= \lim_{\lambda\to\lambda_i} m^\Delta(\lambda) v_i (\delta(\lambda_i))^{-2}\sigma_-, \\ \mathrm{res}(m^\Delta(\lambda);-\lambda_i) &= -\sigma_3 \mathrm{res}(m^\Delta(\lambda);\lambda_i)\sigma_3, \\ \mathrm{res}(m^\Delta(\lambda);\overline{\lambda_i}) &= \lim_{\lambda\to\overline{\lambda_i}} m^\Delta(\lambda)\overline{v_i}(\delta(\overline{\lambda_i}))^2\sigma_+, \\ \mathrm{res}(m^\Delta(\lambda);-\overline{\lambda_i}) &= -\sigma_3 \mathrm{res}(m^\Delta(\lambda);\overline{\lambda_i})\sigma_3, \end{aligned}$$
4. *as* $\lambda \to \infty$, $\lambda \in \mathbb{C}\setminus(\mathcal{Z}_d \cup \widehat{\Gamma})$,
   $$m^\Delta(\lambda) = \mathrm{I} + \mathcal{O}(\lambda^{-1});$$

moreover, $Q(x,t) = 2i\lim_{\lambda\to\infty}(\lambda m^\Delta(x,t;\lambda))_{12}$ *is equal to* $Q(x,t)$ *in Lemma* 2.2, (11).

Proof. Let $m(\lambda)$ be the solution of the RH problem formulated in Lemma 2.1. Define $m^\Delta(\lambda) := m(\lambda)(\Delta(\lambda))^{-1}$. □

In order to motivate Proposition 3.2 and Lemma 3.2 (see below), consider the trajectory of the nth soliton with $\gamma_n \in (\frac{\pi}{2},\pi)$ in the (x,t)-plane which belongs to

the set $\Omega_n := \{(x,t);\, x - 4t\Delta_n^2 \cos\gamma_n = \mathcal{O}(1)\}$, and note from Lemma 2.1 and the soliton ordering in section 2 that, as $t \to +\infty$ and $x \to -\infty$ such that $\lambda_0 > M$ and $(x,t) \in \Omega_n$: (1) $\Re(v_i|_{\Omega_n}) \sim \mathcal{O}(\exp\{-8t\eta_i(\xi_i - \xi_n)\}) \to 0 \,\forall\, i < n$ $(i \in \{1,2,\dots,n-1\})$; (2) $\Re(v_i|_{\Omega_n}) \to \infty \,\forall\, i > n$ $(i \in \{n+1, n+2, \dots, N\})$; and (3) $\Re(v_i|_{\Omega_n}) \sim \mathcal{O}(1)$ for $i = n$. Thus, along the trajectory of the arbitrarily fixed nth soliton, there are exponentially growing polar conditions for solitons i with $n+1 \le i \le N$. One must effectively deal with such growing polar conditions in a self-consistent manner. In a recent paper [33] devoted to the asymptotics of the Toda rarefaction problem, Deift et al. showed how this could be done: they noticed that it is possible to replace the poles with the exponentially growing polar conditions by jump matrices on small, mutually disjoint (*and* disjoint with respect to $\widehat{\Gamma}$) circles such that these jump matrices behave like I + exponentially decreasing terms as $t \to +\infty$. Thus, instead of the original RH problem, one gets a new, "extended" RH problem with $4(N-n)$ fewer poles, and $4(N-n)$ additional circles with jump conditions stated on them. Finally, by removing the added circles from the specification of the extended RH problem, one arrives at the model RH problem: the estimation of the "difference" between the extended and model RH problems shows that the solution of the model RH problem approximates the solution of the original one modulo terms which are decaying exponentially as $t \to +\infty$.

PROPOSITION 3.2. *Introduce arbitrarily small, clockwise and counterclockwise-oriented, mutually disjoint (and disjoint with respect to $\widehat{\Gamma}$) circles $K_j^{\pm}$ and $L_j^{\pm}$, $n+1 \le j \le N$, around the eigenvalues $\{\pm\lambda_j\}_{j=n+1}^N$ and $\{\pm\overline{\lambda_j}\}_{j=n+1}^N$, resp., and define*

$$m^{\flat}(\lambda) := \begin{cases} m^{\Delta}(\lambda), \quad \lambda \in \mathbb{C} \setminus \left(\widehat{\Gamma} \cup \left(\bigcup_{i=n+1}^{N} (K_i^{\pm} \cup L_i^{\pm}) \right) \right), \\ m^{\Delta}(\lambda)\Big(\mathrm{I} - \frac{v_i(\delta(\pm\lambda_i))^{-2}}{(\lambda \mp \lambda_i)} \sigma_- \Big), \quad \lambda \in \mathrm{int} K_i^{\pm}, \quad n+1 \le i \le N, \\ m^{\Delta}(\lambda)\Big(\mathrm{I} + \frac{\overline{v_i}(\delta(\pm\overline{\lambda_i}))^{2}}{(\lambda \mp \overline{\lambda_i})} \sigma_+ \Big), \quad \lambda \in \mathrm{int} L_i^{\pm}, \quad n+1 \le i \le N. \end{cases} \tag{85}$$

Then $m^{\flat}(\lambda)$ solves a RH problem on $(\sigma_{\mathcal{L}} \setminus \cup_{i=n+1}^N (\{\pm\lambda_i\} \cup \{\pm\overline{\lambda_i}\})) \cup (\cup_{i=n+1}^N (K_i^{\pm} \cup L_i^{\pm}))$ with the same jumps as $m^{\Delta}(\lambda)$ on $\widehat{\Gamma}$, $m_+^{\flat}(\lambda) = m_-^{\flat}(\lambda) v^{\Delta}(\lambda)$, and

$$m_+^{\flat}(\lambda) = \begin{cases} m_-^{\flat}(\lambda)\Big(\mathrm{I} + \frac{v_i(\delta(\pm\lambda_i))^{-2}}{(\lambda \mp \lambda_i)} \sigma_- \Big), \quad \lambda \in K_i^{\pm}, \quad n+1 \le i \le N, \\ m_-^{\flat}(\lambda)\Big(\mathrm{I} + \frac{\overline{v_i}(\delta(\pm\overline{\lambda_i}))^{2}}{(\lambda \mp \overline{\lambda_i})} \sigma_+ \Big), \quad \lambda \in L_i^{\pm}, \quad n+1 \le i \le N. \end{cases}$$

Proof. The proof follows from Lemma 3.1 and the definition of $m^{\flat}(\lambda)$. ∎

Remark 3.2. The superscripts $\pm$ on $\{K_i^{\pm}\}_{i=n+1}^N$ and $\{L_i^{\pm}\}_{i=n+1}^N$, which are related to $\{\pm\lambda_i\}_{i=n+1}^N$ and $\{\pm\overline{\lambda_i}\}_{i=n+1}^N$, resp., should *not* be confused with the subscripts $\pm$ appearing in the various RH problems in sections 3–5, namely, $m_{\pm}(\lambda)$, $m_{\pm}^{\Delta}(\lambda)$, $m_{\pm}^{\flat}(\lambda)$, $m_{\pm}^{\sharp}(\lambda)$, $\chi_{\pm}(\lambda)$, $E_{\pm}(\lambda)$, and $\chi_{\pm}^{c}(\lambda)$.

Remark 3.3. Even though the exponentially growing polar (residue) conditions have been replaced by jump matrices, it should be noted that, along the trajectory of soliton n, these jump matrices are also exponentially growing as $t \to +\infty$. These lower/upper diagonal, exponentially growing jump matrices are now replaced, through a sequence of $N-n$ similar transformations, by upper/lower diagonal jump matrices which converge, along the trajectory of soliton n, to I as $t \to +\infty$.

LEMMA 3.2. *Set*

(86)
$$m^\sharp(\lambda):=\begin{cases} m^\flat(\lambda)\prod\limits_{l=n+1}^{N}(d_{l_+}(\lambda))^{-\sigma_3}, & \lambda\in\mathbb{C}\setminus\left(\widehat{\Gamma}\cup\left(\bigcup\limits_{i=n+1}^{N}(K_i^\pm\cup L_i^\pm)\right)\right),\\ m^\flat(\lambda)(J_{K_i^\pm}(\lambda))^{-1}\prod\limits_{l=n+1}^{N}(d_{l_-}(\lambda))^{-\sigma_3}, & \lambda\in\operatorname{int}K_i^\pm,\quad n+1\le i\le N,\\ m^\flat(\lambda)(J_{L_i^\pm}(\lambda))^{-1}\prod\limits_{l=n+1}^{N}(d_{l_-}(\lambda))^{-\sigma_3}, & \lambda\in\operatorname{int}L_i^\pm,\quad n+1\le i\le N,\end{cases}$$

where

(87)
$$d_{l_+}(\lambda):=\tfrac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)},\quad \lambda\in\mathbb{C}\setminus\left(\bigcup_{i=n+1}^{N}(K_i^\pm\cup L_i^\pm)\right),\quad n+1\le l\le N,$$
$$d_{l_-}(\lambda):=\begin{cases}\frac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda\pm\lambda_l)}, & \lambda\in\bigcup\limits_{i=n+1}^{N}\operatorname{int}K_i^\pm,\quad n+1\le l\le N,\\ \frac{(\lambda\pm\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)}, & \lambda\in\bigcup\limits_{i=n+1}^{N}\operatorname{int}L_i^\pm,\quad n+1\le l\le N,\end{cases}$$

and the SL(2,$\mathbb{C}$)*-valued, holomorphic in* $\operatorname{int}K_i^\pm$ *and* $\operatorname{int}L_i^\pm$, *resp., functions* $J_{K_i^\pm}(\lambda)$ *and* $J_{L_i^\pm}(\lambda)$, $n{+}1\le i\le N$, *are given by*

$$J_{K_i^\pm}(\lambda)=\begin{pmatrix}\dfrac{\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}^{-1}(\lambda)}-\frac{v_i(\delta(\pm\lambda_i))^{-2}C_i^\sharp}{(d_{i_-}(\lambda))^2}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}(\lambda)}}{(\lambda\mp\lambda_i)}, & \dfrac{C_i^\sharp}{(d_{i_-}(\lambda))^2}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}(\lambda)}\\ -v_i(\delta(\pm\lambda_i))^{-2}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)}, & (\lambda\mp\lambda_i)\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)}\end{pmatrix},$$
$$J_{L_i^\pm}(\lambda)=\begin{pmatrix}(\lambda\mp\overline{\lambda_i})\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}^{-1}(\lambda)}, & \overline{v_i}(\delta(\pm\overline{\lambda_i}))^2\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}^{-1}(\lambda)}\\ -\frac{\overline{C_i^\sharp}}{(d_{i_-}(\overline{\lambda}))^{-2}}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}^{-1}(\lambda)}, & \dfrac{\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)}-\frac{\overline{v_i}(\delta(\pm\overline{\lambda_i}))^2\overline{C_i^\sharp}}{(d_{i_-}(\lambda))^{-2}}\prod\limits_{l=n+1}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}^{-1}(\lambda)}}{(\lambda\mp\overline{\lambda_i})}\end{pmatrix},$$

with

(88)
$$C_i^\sharp=(v_i)^{-1}(\delta(\pm\lambda_i))^2(d_{i_-}(\pm\lambda_i))^2\prod_{\substack{l=n+1\\ \neq i}}^{N}(d_{l_+}(\pm\lambda_i))^2,\qquad n+1\le i\le N.$$

Then $m^\sharp(\lambda)\colon\mathbb{C}\setminus((\mathcal{Z}_d\setminus\cup_{i=n+1}^{N}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\widehat{\Gamma}\cup(\cup_{i=n+1}^{N}(K_i^\pm\cup L_i^\pm))))\to\mathrm{SL}(2,\mathbb{C})$ *solves the following, extended RH problem on* $(\sigma_{\mathcal{L}}\setminus\cup_{i=n+1}^{N}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\cup_{i=n+1}^{N}(K_i^\pm\cup L_i^\pm))$:

$$m_+^\sharp(\lambda)=m_-^\sharp(\lambda)e^{-i\theta(\lambda)\mathrm{ad}(\sigma_3)}v^\sharp(\lambda),$$

where

$$v^{\sharp}(\lambda)|_{\widehat{\Gamma}}=\begin{pmatrix} (1-r(\lambda)\overline{r(\overline{\lambda})})\frac{\delta_-(\lambda)}{\delta_+(\lambda)}, & \frac{r(\lambda)}{(\delta_-(\lambda)\delta_+(\lambda))^{-1}}\prod_{l=n+1}^{N}\left(\frac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)}\right)^2 \\ -\frac{\overline{r(\overline{\lambda})}}{\delta_-(\lambda)\delta_+(\lambda)}\prod_{l=n+1}^{N}\left(\frac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)}\right)^{-2}, & \frac{\delta_+(\lambda)}{\delta_-(\lambda)} \end{pmatrix},$$

$$v^{\sharp}(\lambda)=\begin{cases} \mathrm{I}+\frac{(v_i)^{-1}(\delta(\pm\lambda_i))^2}{(\lambda\mp\lambda_i)}\left(\frac{\lambda_i^2-\overline{\lambda_i}^2}{2\lambda_i}\right)^2\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\left(\frac{\overline{\lambda_l}^2-\lambda_i^2}{\lambda_l^2-\lambda_i^2}\right)^2\sigma_+, & \lambda\in\bigcup\limits_{i=n+1}^{N}K_i^{\pm}, \\ \mathrm{I}+\frac{(\overline{v_i})^{-1}(\delta(\pm\overline{\lambda_i}))^{-2}}{(\lambda\mp\overline{\lambda_i})}\left(\frac{\lambda_i^2-\overline{\lambda_i}^2}{2\overline{\lambda_i}}\right)^2\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\left(\frac{\lambda_l^2-\overline{\lambda_i}^2}{\overline{\lambda_l}^2-\overline{\lambda_i}^2}\right)^2\sigma_-, & \lambda\in\bigcup\limits_{i=n+1}^{N}L_i^{\pm}, \end{cases} \tag{89}$$

with polar (*residue*) *conditions*,

$$\mathrm{res}(m^{\sharp}(\lambda);\lambda_i)=\lim_{\lambda\to\lambda_i}m^{\sharp}(\lambda)v_i(\delta(\lambda_i))^{-2}\prod_{l=n+1}^{N}\left(\frac{(\lambda_i-\lambda_l)(\lambda_i+\lambda_l)}{(\lambda_i-\overline{\lambda_l})(\lambda_i+\overline{\lambda_l})}\right)^2\sigma_-,\quad 1\le i\le n,$$

$$\mathrm{res}(m^{\sharp}(\lambda);-\lambda_i)=-\sigma_3\mathrm{res}(m^{\sharp}(\lambda);\lambda_i)\sigma_3,\quad 1\le i\le n,$$

$$\mathrm{res}(m^{\sharp}(\lambda);\overline{\lambda_i})=\lim_{\lambda\to\overline{\lambda_i}}m^{\sharp}(\lambda)\overline{v_i}(\delta(\overline{\lambda_i}))^2\prod_{l=n+1}^{N}\left(\frac{(\overline{\lambda_i}-\overline{\lambda_l})(\overline{\lambda_i}+\overline{\lambda_l})}{(\overline{\lambda_i}-\lambda_l)(\overline{\lambda_i}+\lambda_l)}\right)^2\sigma_+,\quad 1\le i\le n,$$

$$\mathrm{res}(m^{\sharp}(\lambda);-\overline{\lambda_i})=-\sigma_3\mathrm{res}(m^{\sharp}(\lambda);\overline{\lambda_i})\sigma_3,\quad 1\le i\le n,$$

and, as $\lambda\to\infty$, $\lambda\in\mathbb{C}\backslash((\mathcal{Z}_d\backslash\cup_{i=n+1}^{N}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\widehat{\Gamma}\cup(\cup_{i=n+1}^{N}(K_i^{\pm}\cup L_i^{\pm}))))$,

$$m^{\sharp}(\lambda)=\mathrm{I}+\mathcal{O}(\lambda^{-1});$$

moreover, $Q(x,t)=2i\lim_{\lambda\to\infty}(\lambda m^{\sharp}(x,t;\lambda))_{12}$ *is equal to* $Q(x,t)$ *in Lemma* 2.2, (11).

Proof. The proof is presented for the eigenvalues $\{\lambda_i\}_{i=n+1}^{N}$, around which are defined the small, clockwise-oriented, mutually disjoint circles $\{K_i^{+}\}_{i=n+1}^{N}$: the proof for the eigenvalues $\{-\lambda_i\}_{i=n+1}^{N}$ and $\{\pm\overline{\lambda_i}\}_{i=n+1}^{N}$ follows in an analogous manner. From the definition of $m^{\sharp}(\lambda)$ and Proposition 3.2, one sees that, on $\{K_i^{+}\}_{i=n+1}^{N}$, $m^{\sharp}(\lambda)$ solves the following RH problem ($\lambda\in\cup_{i=n+1}^{N}K_i^{+}$):

$$m_+^{\sharp}(\lambda)=m_-^{\sharp}(\lambda)\underbrace{\prod_{l=n+1}^{N}(d_{l_-}(\lambda))^{\sigma_3}J_{K_i^{+}}(\lambda)\left(\mathrm{I}+\frac{v_i(\delta(\lambda_i))^{-2}}{(\lambda-\lambda_i)}\sigma_-\right)\prod_{l=n+1}^{N}(d_{l_+}(\lambda))^{-\sigma_3}}_{\text{jump matrix}}.$$

Demanding that the above "jump matrix" be equal to the following upper triangular form, $\mathrm{I}+\frac{C_i^{\sharp}}{(\lambda-\lambda_i)}\sigma_+$, $n+1\le i\le N$, one shows that

$$J_{K_i^{+}}(\lambda)=\begin{pmatrix} \frac{d_{i_+}(\lambda)}{d_{i_-}(\lambda)}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_+}(\lambda)}{d_{l_-}(\lambda)}-\frac{v_i(\delta(\lambda_i))^{-2}C_i^{\sharp}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}(\lambda)}}{(\lambda-\lambda_i)^2d_{i_-}(\lambda)d_{i_+}(\lambda)}, & \frac{C_i^{\sharp}}{(\lambda-\lambda_i)}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}(\lambda)}\frac{}{d_{i_-}(\lambda)d_{i_+}(\lambda)} \\ -\frac{v_i(\delta(\lambda_i))^{-2}}{(\lambda-\lambda_i),}\frac{d_{i_-}(\lambda)}{d_{i_+}(\lambda)}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)}, & \frac{d_{i_-}(\lambda)}{d_{i_+}(\lambda)}\prod\limits_{\substack{l=n+1\\ \neq i}}^{N}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)} \end{pmatrix}.$$

Note that $\det(J_{K_i^+}(\lambda))=1$ $(n+1\le i\le N)$. Defining, for $n+1\le l\le N$, $d_{l_+}(\lambda)$ and $d_{l_-}(\lambda)$ as in (87), and choosing $C_i^\sharp$, $n+1\le i\le N$, as in (88) (with $+\lambda_i$), one gets the expression for $J_{K_i^+}(\lambda)$ (which is holomorphic for all $\lambda\in\cup_{i=n+1}^N \mathrm{int}K_i^+$) given in the lemma; also, because of the symmetry properties of $\delta(\lambda)$ (Proposition 3.1), $\overline{C_i^\sharp}=(\overline{v_i})^{-1}(\delta(\pm\overline{\lambda_i}))^{-2}\cdot(d_{i_-}(\pm\overline{\lambda_i}))^{-2}\prod_{\substack{l=n+1\\ \neq i}}^N(d_{l_+}(\pm\overline{\lambda_i}))^{-2}$. The remainder of the proof is a consequence of Lemma 3.1, Proposition 3.2, and the definition of $m^\sharp(\lambda)$. □

Remark 3.4. Even though, along the trajectory of soliton n, all the initial, exponentially growing nilpotent residue matrices have been replaced by jump matrices which tend to I as $t\to+\infty$, i.e., $\exists\,\varepsilon\in\mathbb{R}_{>0}$ such that $\forall\, i\in\{n+1,n+2,\dots,N\}$ $|(v_i|_{\Omega_n})^{-1}|\sim\mathcal{O}(\exp\{-\varepsilon t\})$, it does not necessarily follow that elements in the solution of the extended RH problem for $m^\sharp(\lambda)$ cannot grow exponentially; for example, note that the $(2\,1)$-elements of $J_{K_i^\pm}(\lambda)$ and the $(1\,2)$-elements of $J_{L_i^\pm}(\lambda)$, $n+1\le i\le N$, grow exponentially.

By estimating the error, along the trajectory of soliton n $(1\le n\le N)$, when the jump matrices on $\{K_i^\pm,L_i^\pm\}_{i=n+1}^N$ are removed from the specification of the RH problem for $m^\sharp(\lambda)$, one gets the following—asymptotically solvable—model RH problem.

LEMMA 3.3. *Let* $\chi(\lambda)$ *solve the following RH problem on* $\sigma_{\mathcal{L}}\setminus\cup_{i=n+1}^N(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})$:

$$\chi_+(\lambda)=\chi_-(\lambda)e^{-i\theta(\lambda)\mathrm{ad}(\sigma_3)}v^\sharp(\lambda)|_{\widehat{\Gamma}},\qquad \lambda\in\widehat{\Gamma},$$

with polar (*residue*) *conditions*

$$\begin{aligned}
\mathrm{res}(\chi(\lambda);\lambda_i)&=\lim_{\lambda\to\lambda_i}\chi(\lambda)v_i(\delta(\lambda_i))^{-2}\prod_{l=n+1}^N\left(\frac{(\lambda_i-\lambda_l)(\lambda_i+\lambda_l)}{(\lambda_i-\overline{\lambda_l})(\lambda_i+\overline{\lambda_l})}\right)^2\sigma_-,\qquad 1\le i\le n,\\
\mathrm{res}(\chi(\lambda);-\lambda_i)&=-\sigma_3\mathrm{res}(\chi(\lambda);\lambda_i)\sigma_3,\qquad 1\le i\le n,\\
\mathrm{res}(\chi(\lambda);\overline{\lambda_i})&=\lim_{\lambda\to\overline{\lambda_i}}\chi(\lambda)\overline{v_i}(\delta(\overline{\lambda_i}))^{2}\prod_{l=n+1}^N\left(\frac{(\overline{\lambda_i}-\overline{\lambda_l})(\overline{\lambda_i}+\overline{\lambda_l})}{(\overline{\lambda_i}-\lambda_l)(\overline{\lambda_i}+\lambda_l)}\right)^2\sigma_+,\qquad 1\le i\le n,\\
\mathrm{res}(\chi(\lambda);-\overline{\lambda_i})&=-\sigma_3\mathrm{res}(\chi(\lambda);\overline{\lambda_i})\sigma_3,\qquad 1\le i\le n,
\end{aligned}$$

and, as $\lambda\to\infty$, $\lambda\in\mathbb{C}\setminus(\cup_{i=1}^n(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})\cup\widehat{\Gamma})$,

$$\chi(\lambda)=\mathrm{I}+\mathcal{O}(\lambda^{-1}).$$

Then as $t\to+\infty$ *and* $x\to-\infty$ *such that* $\lambda_0>M$, *and* $(x,t)\in\Omega_n$, *the function* $E(\lambda):=m^\sharp(\lambda)(\chi(\lambda))^{-1}$ *has the following asymptotics:*

$$E(\lambda)=\mathrm{I}+\mathcal{O}(F(\lambda;\lambda_0)\exp\{-abt\}),\tag{90}$$

where $||F(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C};M_2(\mathbb{C}))}<\infty$, $||F(\lambda;\cdot)||_{\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{C}))}<\infty$, $F(\lambda;\lambda_0)\sim\mathcal{O}\left(\frac{C(\lambda_0)}{\lambda}\right)$ *as* $\lambda\to\infty$ *with* $C(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{C}))$, $a:=8\min\{\eta_i\}_{i=n+1}^N\,(>0)$, *and* $b:=\min\{|\xi_n-\xi_i|\}_{i=n+1}^N$.

Proof. Writing, for $n+1\le i\le N$, (89) in the following form:

$$v^\sharp(\lambda):=\begin{cases}\mathrm{I}+(v_i)^{-1}\widetilde{\mathcal{W}}_{K_i^\pm}(\lambda)\sigma_+, & \lambda\in\bigcup_{i=n+1}^N K_i^\pm,\\ \mathrm{I}+(\overline{v_i})^{-1}\widetilde{\mathcal{W}}_{L_i^\pm}(\lambda)\sigma_-, & \lambda\in\bigcup_{i=n+1}^N L_i^\pm,\end{cases}$$

consider the "error function" $E(\lambda)$ defined in the lemma. One notes that (1) $\det(E(\lambda)) = 1$; (2) $E(\lambda)$ has no poles; and (3) $E(\lambda)$ solves the following RH problem on the oriented contour $\Sigma_E := \cup_{i=n+1}^{N}(K_i^+ \cup K_i^- \cup L_i^+ \cup L_i^-)$:

$$E_+(\lambda) = E_-(\lambda)\left(\mathrm{I} + (v_i)^{-1}\widetilde{\mathcal{W}}_{K_i^{\pm}}(\lambda)\begin{pmatrix} -\chi_{11}(\lambda)\chi_{21}(\lambda) & (\chi_{11}(\lambda))^2 \\ -(\chi_{21}(\lambda))^2 & \chi_{11}(\lambda)\chi_{21}(\lambda) \end{pmatrix}\right), \quad \lambda \in K_i^{\pm},$$

$$E_+(\lambda) = E_-(\lambda)\left(\mathrm{I} + (\overline{v_i})^{-1}\widetilde{\mathcal{W}}_{L_i^{\pm}}(\lambda)\begin{pmatrix} \chi_{12}(\lambda)\chi_{22}(\lambda) & -(\chi_{12}(\lambda))^2 \\ (\chi_{22}(\lambda))^2 & -\chi_{12}(\lambda)\chi_{22}(\lambda) \end{pmatrix}\right), \quad \lambda \in L_i^{\pm},$$

$n+1 \le i \le N$, and, as $\lambda \to \infty$, $\lambda \in \mathbb{C} \setminus \Sigma_E$, $E(\lambda) = \mathrm{I} + \mathcal{O}(\lambda^{-1})$. Now, writing the RH problem for $E(\lambda)$ on the oriented contour Σ_E in terms of an equivalent system of linear singular integral equations, using the explicit asymptotic solution of the model RH problem for $\chi(\lambda)$ given in section 4, recalling that, as $t \to +\infty$ and $x \to -\infty$ such that $\lambda_0 > M$ and $(x,t) \in \Omega_n$, $(v_i|_{\Omega_n})^{-1} \sim \mathcal{O}(\exp\{-8t\eta_i|\xi_n - \xi_i|\})$, $n+1 \le i \le N$, and proceeding as in the proof of Lemma 3.3 in [33], one deduces the estimate in (90). □

4. Asymptotic solution of the model RH problem. In this section, the asymptotic (as $t \to +\infty$ and $x/t \sim \mathcal{O}(1)$) solution of the model RH problem (Lemma 3.3) for the Schwartz class of nonreflectionless generic potentials ($r(\lambda) \in \mathcal{S}(\widehat{\Gamma};\mathbb{C})$) is presented. Before doing so, however, recall the following well-known fact from the matrix RH theory [22, 24].

PROPOSITION 4.1. *The solution of the model RH problem* (*Lemma* 3.3), $\chi(\lambda)\colon \mathbb{C} \setminus (\widehat{\Gamma} \cup (\cup_{i=1}^{n}(\{\pm\lambda_i\} \cup \{\pm\overline{\lambda_i}\}))) \to \mathrm{SL}(2,\mathbb{C})$, *has the following representation:*

$$\chi(\lambda) = \chi_d(\lambda) + \int_{\widehat{\Gamma}} \frac{\chi_-(\varrho)(v^{\sharp}(\varrho)|_{\widehat{\Gamma}} - \mathrm{I})}{(\varrho - \lambda)} \frac{d\varrho}{2\pi i}, \tag{91}$$

where

$$\chi_d(\lambda) = \mathrm{I} + \sum_{i=1}^{n}\left(\frac{\mathrm{res}(\chi(\lambda);\lambda_i)}{(\lambda - \lambda_i)} - \frac{\sigma_3 \mathrm{res}(\chi(\lambda);\lambda_i)\sigma_3}{(\lambda + \lambda_i)} + \frac{\mathrm{res}(\chi(\lambda);\overline{\lambda_i})}{(\lambda - \overline{\lambda_i})} - \frac{\sigma_3 \mathrm{res}(\chi(\lambda);\overline{\lambda_i})\sigma_3}{(\lambda + \overline{\lambda_i})}\right). \tag{92}$$

The solution of (91) *can be written as the following ordered product:*

$$\chi(\lambda) = \chi_d(\lambda)\chi^c(\lambda), \tag{93}$$

where $\chi_d(\lambda)$ *is given by* (92), *and* $\chi^c(\lambda)$ *solves the following RH problem:* (1) $\chi^c(\lambda)$ *is piecewise holomorphic for all* $\lambda \in \mathbb{C} \setminus \widehat{\Gamma}$; (2) $\chi_+^c(\lambda) = \chi_-^c(\lambda)\exp\{-i\theta(\lambda)\mathrm{ad}(\sigma_3)\}(v^{\sharp}(\lambda)|_{\widehat{\Gamma}})$, $\lambda \in \widehat{\Gamma}$; *and* (3) *as* $\lambda \to \infty$, $\lambda \in \mathbb{C} \setminus \widehat{\Gamma}$, $\chi^c(\lambda) = \mathrm{I} + \mathcal{O}(\lambda^{-1})$.

Remark 4.1. From Proposition 4.1, (93), it is seen that, in order to solve the model RH problem, explicit knowledge of $\chi_d(\lambda)$ and $\chi^c(\lambda)$ is necessary. The determination of $\chi^c(\lambda)$ is technically the more complicated of the two: actually, the determination of $\chi_d(\lambda)$ depends on the explicit knowledge of $\chi^c(\lambda)$ (see Proposition 4.2); hence, the asymptotic solution of $\chi^c(\lambda)$ is presented first (see Lemma 4.1).

In order to more fully comprehend certain elements of the proof of Lemma 4.1 given below, the Beals–Coifman [24, 25] formulation for the solution of a (matrix) RH problem on an oriented contour is requisite: a self-contained synopsis of this

formulation as it applies to the solution of the RH problem for $\chi^c(\lambda)$ stated in Proposition 4.1 now follows. Writing the jump matrix in the following factorized form: $v^\sharp(\lambda)|_{\widehat{\Gamma}} := (\mathrm{I}-w^-_{x,t}(\lambda))^{-1}(\mathrm{I}+w^+_{x,t}(\lambda))$, $\lambda\in\widehat{\Gamma}$, where $w^\pm_{x,t}(\lambda)\in\cap_{k\in\{2,\infty\}}\mathcal{L}^k(\widehat{\Gamma};M_2(\mathbb{C}))$ (with $||w^\pm_{x,t}(\cdot)||_{\cap_{k\in\{2,\infty\}}\mathcal{L}^k(\widehat{\Gamma};M_2(\mathbb{C}))} := \sum_{k\in\{2,\infty\}}||w^\pm_{x,t}(\cdot)||_{\mathcal{L}^k(\widehat{\Gamma};M_2(\mathbb{C}))}$), resp., are nilpotent off-diagonal upper/lower triangular matrices, define $w_{x,t}(\lambda) := w^-_{x,t}(\lambda) + w^+_{x,t}(\lambda)$, and introduce the operator $C_{w_{x,t}}$ on $\mathcal{L}^2(\widehat{\Gamma};M_2(\mathbb{C}))$ as $C_{w_{x,t}}f := C_+(f\,w^-_{x,t}) + C_-(f\,w^+_{x,t})$, where $f\in\mathcal{L}^2(\widehat{\Gamma};M_2(\mathbb{C}))$, and $C_\pm\colon\mathcal{L}^2(\widehat{\Gamma};M_2(\mathbb{C}))\to\mathcal{L}^2(\widehat{\Gamma};M_2(\mathbb{C}))$ denote the Cauchy operators

$$(C_\pm f)(\lambda) := \lim_{\substack{\lambda'\to\lambda\\ \lambda'\in\pm\text{ side of }\widehat{\Gamma}}}\int_{\widehat{\Gamma}}\frac{f(\varrho)}{(\varrho-\lambda')}\frac{d\varrho}{2\pi i}.$$

THEOREM 4.1 (see [24]). *If $\mu^c(\lambda)\in\mathrm{I}\oplus\mathcal{L}^2(\widehat{\Gamma};M_2(\mathbb{C}))$ solves the following linear singular integral equation:*

$$(\underline{\mathbf{Id}} - C_{w_{x,t}})\mu^c = \mathrm{I},$$

where $\underline{\mathbf{Id}}$ is the identity operator on $\mathrm{I}\oplus\mathcal{L}^2(\widehat{\Gamma};M_2(\mathbb{C}))$, then the solution of the RH problem for $\chi^c(\lambda)$ is

$$\chi^c(\lambda) = \mathrm{I} + \int_{\widehat{\Gamma}}\frac{\mu^c(\varrho)w_{x,t}(\varrho)}{(\varrho-\lambda)}\frac{d\varrho}{2\pi i},\qquad \lambda\in\mathbb{C}\backslash\widehat{\Gamma},$$

where $\mu^c(\lambda) = \chi^c_+(\lambda)(\mathrm{I}+w^+_{x,t}(\lambda))^{-1} = \chi^c_-(\lambda)(\mathrm{I}-w^-_{x,t}(\lambda))^{-1}$.

LEMMA 4.1. *Let ϵ_0 denote an arbitrarily fixed, sufficiently small positive real number. For $\aleph\in\{0,\pm\lambda_0\}$, set $\mathcal{N}(\aleph;\epsilon_0) := \{\lambda;\,|\lambda-\aleph|\le\epsilon_0\}$. Then as $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0 > M$ and $\lambda\in\mathbb{C}\backslash\cup_{\aleph\in\{0,\pm\lambda_0\}}\mathcal{N}(\aleph;\epsilon_0)$, $\chi^c(\lambda)$ has the following asymptotic expansion:*

$$\chi^c(\lambda) = \mathrm{I} + \frac{1}{4}\sqrt{\frac{\nu(\lambda_0)}{2\lambda_0^2 t}}\left(\frac{1}{\lambda-\lambda_0}+\frac{1}{\lambda+\lambda_0}\right)\Big(\exp\{-i(\phi^+(\lambda_0)+\widehat{\Phi}^+(\lambda_0;t))\}\sigma_- + \exp\{i(\phi^+(\lambda_0)+\widehat{\Phi}^+(\lambda_0;t))\}\sigma_+\Big) + \mathcal{O}\left(\frac{G(\lambda;\lambda_0)\ln t}{t}\right),$$

where $\nu(\lambda_0)$, $\phi^+(\lambda_0)$, and $\widehat{\Phi}^+(\lambda_0;t)$ are given in Theorem 2.1, equations (21), (22), and (24), $||G(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C}\backslash\cup_{\aleph\in\{0,\pm\lambda_0\}}\mathcal{N}(\aleph;\epsilon_0);M_2(\mathbb{C}))}<\infty$, $G(\lambda;\cdot)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{C}))$, $G(\lambda;\lambda_0)\sim\mathcal{O}(\frac{C(\lambda_0)}{\lambda})$ as $\lambda\to\infty$ with $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{C}))$, and satisfies the following involutions: $\chi^c(-\lambda) = \sigma_3\chi^c(\lambda)\sigma_3$ and $\chi^c(\lambda) = \sigma_1\overline{\chi^c(\overline{\lambda})}\sigma_1$.

Proof. In sections 5 and 6 of [15], it was shown that, for $\lambda\in\mathbb{C}\backslash\cup_{\aleph\in\{0,\pm\lambda_0\}}\mathcal{N}(\aleph;\epsilon_0)$, as $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0 > M$, for arbitrary $l'\in\mathbb{Z}_{\ge1}$,

$$\chi^c(\lambda) = \mathrm{I} + \sum_{\vec{m}_3\in\mathbb{M}_3}\int_{a_l(\vec{m}_3)}^{a_u(\vec{m}_3)}\frac{m_3\mu^c(\varsigma)|_{\mathrm{sgn}(m_3)}(\delta(\varsigma))^{2m_3}e^{-2im_3t\theta(\varsigma)}\mathcal{R}^{\vec{m}_3}(\varsigma)\sigma_{\mathrm{sgn}(m_3)}}{(\varsigma-\lambda)}\frac{d\varsigma}{2\pi i} + \sum_{\vec{m}_2\in\mathbb{M}_2}\int_0^{b(\vec{m}_2)}\frac{m_2\mu^c(\varsigma)|_{\mathrm{sgn}(m_2)}(\delta(\varsigma))^{2m_2}e^{-2im_2t\theta(\varsigma)}\mathcal{R}^{\vec{m}_2}(\varsigma)\sigma_{\mathrm{sgn}(m_2)}}{(\varsigma-\lambda)}\frac{d\varsigma}{2\pi i} + \mathcal{O}\left(\frac{\underline{c}(\lambda_0)}{(\lambda_0^2t)^{l'}}\right),$$

where (1) $\vec{m}_k \in \mathbb{M}_k$ denotes the set of vectors with k components, each of which take the values ± 1 ($\text{card}(\mathbb{M}_k) = 2^k$); (2) $a_l(\vec{m}_3) = m_1\lambda_0 + m_2\varepsilon \exp\{-\frac{i\pi m_3}{4}\}$, $a_u(\vec{m}_3) = m_1\lambda_0$, and $b(\vec{m}_2) = m_1\varepsilon \exp\{\frac{i\pi m_2}{4}\}$, where ε is an arbitrarily fixed, sufficiently small positive real number; (3) $\mu^c(\varsigma)|_+ := \mu^c(\varsigma)|_{L'}$, $\mu^c(\varsigma)|_- := \mu^c(\varsigma)|_{\overline{L'}}$, with $L' = \{\lambda;\ \lambda = \widehat{u}\exp\{\frac{i\pi}{4}\},\ \widehat{u} \in (-\varepsilon, \varepsilon)\} \cup (\cup_{l\in\{\pm 1\}}\{\lambda;\ \lambda = l\lambda_0 + \widehat{u}\exp\{-\frac{i\pi}{4}\},\ \widehat{u} \in (-\varepsilon, \varepsilon)\})$, and where $\mu^c(\cdot)$ is the solution of the Beals–Coifman [24] linear singular integral equation (Theorem 4.1); (4) $\theta(\varsigma) = 2\varsigma^2(\varsigma^2 - 2\lambda_0^2)$; (5) $\delta(\varsigma) = ((\frac{\varsigma-\lambda_0}{\varsigma})(\frac{\varsigma+\lambda_0}{\varsigma}))^{i\nu} \exp\{\sum_{l\in\{\pm\}}(\rho_l(\varsigma) + \widehat{\rho}_l(\varsigma))\}$, $\nu := \nu(\lambda_0)$, $\rho_\pm(\varsigma) = \frac{1}{2\pi i}\int_0^{\pm\lambda_0} \ln\left(\frac{1-|r(\varrho)|^2}{1-|r(\lambda_0)|^2}\right)\frac{d\varrho}{(\varrho-\varsigma)}$, and $\widehat{\rho}_\pm(\varsigma) = \int_{\pm i\infty}^{i0} \frac{\ln(1-r(\varrho)\overline{r(\overline{\varrho})})}{(\varrho-\varsigma)}\frac{d\varrho}{2\pi i}$; (6) $\mathcal{R}^{-1,-1,1}(\varsigma) = \mathcal{R}^{1,1,1}(\varsigma) = -r(\varsigma)\mathcal{P}(\varsigma)$, $\mathcal{R}^{-1,1,1}(\varsigma) = \mathcal{R}^{1,-1,1}(\varsigma) = \frac{r(\varsigma)\mathcal{P}(\varsigma)}{(1-r(\varsigma)\overline{r(\overline{\varsigma})})}$, $\mathcal{R}^{-1,-1,-1}(\varsigma) = \mathcal{R}^{1,1,-1}(\varsigma) = (\mathcal{R}^{1,1,1}(\varsigma))^*$, and $\mathcal{R}^{-1,1,-1}(\varsigma) = \mathcal{R}^{1,-1,-1}(\varsigma) = (\mathcal{R}^{1,-1,1}(\varsigma))^*$, where $\mathcal{P}(z) := \prod_{l=n+1}^{N}\left(\frac{(z-\overline{\lambda_l})(z+\overline{\lambda_l})}{(z-\lambda_l)(z+\lambda_l)}\right)^2$, and $\alpha(\cdot) = (\beta(\cdot))^*$ means that $\alpha(\cdot)$ is the same piecewise-rational function as $\beta(\cdot)$ except with the complex conjugated coefficients; (7) $\mathcal{R}^{1,1}(\varsigma) = \mathcal{R}^{-1,1}(\varsigma) = \frac{r(\varsigma)\mathcal{P}(\varsigma)}{(1-|r(\varsigma)|^2)} - \frac{r(i\varsigma)\mathcal{P}(i\varsigma)}{(1+|r(i\varsigma)|^2)}$ and $\mathcal{R}^{1,-1}(\varsigma) = \mathcal{R}^{-1,-1}(\varsigma) = -(\mathcal{R}^{1,1}(\varsigma))^*$; and (8) $\underline{c}(\lambda_0) \in \mathcal{L}^\infty(\mathbb{R}_{>M}; M_2(\mathbb{C}))$. Since, in the above expression for $\chi^c(\lambda)$, the estimation of all the integrals is analogous, without loss of generality, the following integral is considered:

$$I_0 := \int_{\lambda_0 + \varepsilon e^{-\frac{i\pi}{4}}}^{\lambda_0} \mathcal{A}_0(\varsigma)\mathcal{B}_0(\varsigma)\frac{d\varsigma}{2\pi i},$$

where $\mathcal{A}_0(\varsigma) := \mu^c(\varsigma)|_{L'}$, and $\mathcal{B}_0(\varsigma) := -\frac{\delta^2(\varsigma)\exp\{-2it\theta(\varsigma)\}r(\varsigma)\mathcal{P}(\varsigma)}{(\varsigma-\lambda)}\sigma_+$. Begin by estimating $\mathcal{B}_0(\varsigma)\frac{d\varsigma}{2\pi i}$: (1) expand $\mathcal{B}_0(\varsigma)$ in a Taylor series about λ_0; (2) make the following change of variable [15]: $\varsigma := \varsigma(\widetilde{w}) = \lambda_0 + \widetilde{w}(16\lambda_0^2 t)^{-1/2}$, and express the expansion obtained in (1) above in terms of $\widetilde{w}$; and (3) use the following identity: $ab = (a-1)(b-1) + (a-1) + (b-1) + 1$. Carrying out steps (1)–(3), one gets that

$$\begin{aligned}
\mathcal{B}_0(\varsigma)\frac{d\varsigma}{2\pi i}|_{\varsigma(\widetilde{w})} = &-\frac{(\widetilde{w})^{2i\nu}\lambda_0^{-4i\nu}t^{-i\nu}2^{-2i\nu}e^{2s(\lambda_0)}e^{4i\lambda_0^4 t}e^{-i\widetilde{w}^2}d\widetilde{w}\sigma_+}{(\lambda-\lambda_0)(2\pi i)\sqrt{16\lambda_0^2 t}}\left\{\mathcal{R}(\lambda_0)\right.\\
&\left.+\left(\mathcal{R}'(\lambda_0) - \frac{3i\nu\mathcal{R}(\lambda_0)}{\lambda_0}\right)\frac{\widetilde{w}}{\sqrt{16\lambda_0^2 t}}\right\}\sum_{\vec{l}_4\in\mathbb{L}_4}\prod_{k=1}^{4}(p_k(\widetilde{w}))^{l_k}\\
&-\frac{(\widetilde{w})^{2i\nu+1}\lambda_0^{-4i\nu}t^{-i\nu}2^{-2i\nu}e^{2s(\lambda_0)}e^{4i\lambda_0^4 t}e^{-i\widetilde{w}^2}d\widetilde{w}\sigma_+}{(\lambda-\lambda_0)^2(2\pi i)(16\lambda_0^2 t)}\\
\times\mathcal{R}(\lambda_0)\sum_{\vec{l}_4\in\mathbb{L}_4}\prod_{k=1}^{4}(p_k(\widetilde{w}))^{l_k} &+ \mathcal{O}\left(\left\{\frac{C_1^\flat(\lambda_0)}{(\lambda-\lambda_0)} + \frac{C_2^\flat(\lambda_0)}{(\lambda-\lambda_0)^2}\right\}\frac{e^{4i\lambda_0^4 t}\widetilde{w}^{2i\nu+2}e^{-i\widetilde{w}^2}d\widetilde{w}\sigma_+}{t^{3/2+i\nu}}\right),
\end{aligned}$$

where $\vec{l}_k \in \mathbb{L}_k$ denotes the set of vectors with k components, each of which take the values 0 and 1 ($\text{card}(\mathbb{L}_k) = 2^k$), $s(\lambda_0) := \sum_{l\in\{\pm\}}(\rho_l(\lambda_0) + \widehat{\rho}_l(\lambda_0))$, $\mathcal{R}(\lambda_0) := -r(\lambda_0)\mathcal{P}(\lambda_0)$, $\mathcal{R}'(\lambda_0) = -r'(\lambda_0)\mathcal{P}(\lambda_0) - r(\lambda_0)\mathcal{P}(\lambda_0)\sum_{k=n+1}^{N}\{\frac{4i\sin(\arg(\lambda_0-\lambda_k))}{|\lambda_0-\lambda_k|} + \frac{4i\sin(\arg(\lambda_0+\lambda_k))}{|\lambda_0+\lambda_k|}\}$, $(\bullet)'(\lambda_0) := \frac{d(\bullet)(z)}{dz}|_{z=\lambda_0}$,

$$p_k(\widetilde{w}) := \exp\left\{2\widetilde{\Delta}_k\left(\lambda_0 + \frac{\widetilde{w}}{\sqrt{16\lambda_0^2 t}}\right)\right\} - 1, \qquad k \in \{1, 2\},$$

$$p_3(\widetilde{w}) := \exp\left\{-\frac{i\widetilde{w}^3}{\lambda_0\sqrt{16\lambda_0^2 t}}\right\} - 1, \quad p_4(\widetilde{w}) := \exp\left\{-\frac{i\widetilde{w}^4}{4^3\lambda_0^4 t}\right\} - 1,$$

$$\widetilde{\Delta}_1\left(\lambda_0 + \frac{\widetilde{w}}{\sqrt{16\lambda_0^2 t}}\right) = \sum_{l\in\{\pm\}}\left(\rho_l\left(\lambda_0 + \frac{\widetilde{w}}{\sqrt{16\lambda_0^2 t}}\right) - \rho_l(\lambda_0)\right),$$

$$\widetilde{\Delta}_2\left(\lambda_0 + \frac{\widetilde{w}}{\sqrt{16\lambda_0^2 t}}\right) = \sum_{l\in\{\pm\}}\left(\widehat{\rho}_l\left(\lambda_0 + \frac{\widetilde{w}}{\sqrt{16\lambda_0^2 t}}\right) - \widehat{\rho}_l(\lambda_0)\right),$$

and $C_i^{\flat}(\lambda_0) \in \mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$, $i \in \{1,2\}$. Now, proceed to estimate $\mu^c(\varsigma)|_{L'}$ for $\varsigma \in (\lambda_0, \lambda_0 + \varepsilon\exp\{-\frac{i\pi}{4}\})$: for this, the Beals–Coifman [24] formulation for the solution of a RH problem on an oriented contour is necessary (Theorem 4.1 and the paragraph preceding it for discussion and notation); in particular, one has to estimate the functions $w_{x,t}^{\pm}(\varsigma)$ on $L' \cup \overline{L'}$. In section 5 of [15], it was shown that, on $(\lambda_0, \lambda_0 + \varepsilon\exp\{-\frac{i\pi}{4}\})$, $w_{x,t}^{+}(\varsigma) = -(\delta(\varsigma))^{\mathrm{ad}(\sigma_3)}\exp\{-it\theta(\varsigma)\mathrm{ad}(\sigma_3)\}r(\varsigma)\mathcal{P}(\varsigma)\sigma_+$ and $w_{x,t}^{-}(\varsigma) = 0$; hence, from the Beals–Coifman [24] formulation, for any $f \in \mathcal{L}^2(L';M_2(\mathbb{C}))$, $C_{w_{x,t}}f = C_-(f\, w_{x,t}^{+})$. To estimate $w_{x,t}^{+}(\varsigma)$, one proceeds as follows: (1) recalling that $r(\varsigma) \in \mathcal{S}(\widehat{\Gamma};\mathbb{C})$ and $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})} < 1$, expand $(\delta(\varsigma))^{\mathrm{ad}(\sigma_3)}$, for $\varsigma \in (\lambda_0, \lambda_0 + \varepsilon\exp\{-\frac{i\pi}{4}\})$, via an integration by parts argument; (2) expand $\exp\{-it\theta(\varsigma)\mathrm{ad}(\sigma_3)\}r(\varsigma)\mathcal{P}(\varsigma)$ in a Taylor series about λ_0; and (3) change variables [15], $\varsigma := \varsigma(\widetilde{w})$. Carrying out steps (1)–(3), one shows that

$$w_{x,t}^{+}(\varsigma)\big|_{\varsigma(\widetilde{w})} = e^{i\{4\lambda_0^4 t - \widetilde{w}^2 + 2\nu\ln(\widetilde{w}/2) - \nu\ln t + 2\phi(\lambda_0)\}}\left(v_{00}(\widetilde{w};\lambda_0) + \frac{v_{10}(\widetilde{w};\lambda_0) + v_{11}(\widetilde{w};\lambda_0)\ln t}{\sqrt{t}}\right) + \mathcal{O}\left(\frac{v_{22}(\widetilde{w};\lambda_0)(\ln t)^2}{t}\right),$$

where $\phi(\lambda_0) = 2\nu\ln\lambda_0 + \frac{1}{2\pi}\int_0^{\lambda_0}\ln|z^2 - \lambda_0^2|d\ln(1-|r(z)|^2) - \frac{1}{2\pi}\int_0^{\infty}\ln|z^2 + \lambda_0^2|d\ln(1 + |r(iz)|^2)$, $v_{00}(\widetilde{w};\lambda_0) := \mathcal{R}(\lambda_0)$, $v_{10}(\widetilde{w};\lambda_0), v_{11}(\widetilde{w};\lambda_0)$, and $v_{22}(\widetilde{w};\lambda_0)$ are nilpotent matrix polynomials whose elements are sums of products of terms of the type $\widetilde{w}^j$ and $(\ln\widetilde{w})^k$, $j \in \mathbb{Z}_{\geq 1}$, $k \in \mathbb{Z}_{\geq 0}$, with λ_0-dependent coefficients, and, for $0 \leq j \leq 2$, $0 \leq k \leq j$,

$$||\exp\{-i(\cdot)^2\}\exp\{2i\nu\ln(\cdot)\}v_{jk}(\cdot;\lambda_0)||_{\cap_{l\in\{1,2,\infty\}}\mathcal{L}^l(L_s'\setminus\{0\};\mathbb{C})} < \infty,$$

with $||(\cdot)||_{\cap_{l\in\{1,2,\infty\}}\mathcal{L}^l(L_s'\setminus\{0\};\mathbb{C})} := \sum_{l\in\{1,2,\infty\}}||(\cdot)||_{\mathcal{L}^l(L_s'\setminus\{0\};\mathbb{C})}$, where $L_s'\setminus\{0\}$ denotes the scaled and shifted version of $L'\setminus\{\lambda_0\}$. Hence, for any $f \in \mathcal{L}^l(L_s';M_2(\mathbb{C}))$, $l \in \{2,\infty\}$,

$$(C_{w_{x,t}}f)(\widetilde{w};\lambda_0) = \sum_{j=0}^{1}\sum_{k=0}^{j}\frac{(\ln t)^k}{t^{j/2}}(C_{jk}^{\natural}f)(\widetilde{w};\lambda_0) + \mathcal{O}\left(\frac{(\ln t)^2}{t}(C_{22}^{\natural}f)(\widetilde{w};\lambda_0)\right),$$

where

$$\frac{\exp\{2i\nu\ln 2\}}{\exp\{2i\phi(\lambda_0)\}}(C_{jk}^{\natural}f)(\widetilde{w};\lambda_0) := \lim_{\substack{\lambda'\to\widetilde{w}\\ \lambda'\in-\text{ side of }L_s'}}\int_{L_s'}\frac{f(z)\exp\{-iz^2\}z^{2i\nu}v_{jk}(z;\lambda_0)}{(z-\lambda')}\frac{dz}{2\pi i},$$

$$0 \leq j \leq 2, 0 \leq k \leq j,$$

and $||C_{jk}^{\natural}(\cdot;\lambda_0)||_{\mathcal{M}(L_s'\setminus\{0\};M_2(\mathbb{C}))} \leq \mathcal{K}_1^{\natural}(\lambda_0) < \infty$, with $\mathcal{M}(\bullet;M_2(\mathbb{C}))$ denoting the space of bounded linear operators acting from $\mathcal{L}^l(\bullet;M_2(\mathbb{C}))$ into $\mathcal{L}^2(\bullet;M_2(\mathbb{C}))$, $l \in \{2,\infty\}$.

According to Theorem 4.1, $\mu^c(\cdot)$ satisfies the following linear singular integral equation on $L' \cup \overline{L'}$, $(\underline{\mathbf{Id}} - C_{w_{x,t}})\mu^c = \mathrm{I}$; hence, $\mu^c = (\underline{\mathbf{Id}} - C_{w_{x,t}})^{-1}\mathrm{I}$. It was shown in [15] that, as $t \to +\infty$ and $x \to -\infty$ such that $\lambda_0 > M$, $\ker(\underline{\mathbf{Id}} - C_{w_{x,t}}) = \emptyset$ and $||(\underline{\mathbf{Id}} - C^{\natural}_{00}(\cdot;\lambda_0))^{-1}||_{\mathcal{M}(L'_s\setminus\{0\};M_2(\mathbb{C}))} \le \mathcal{K}^{\natural}_2(\lambda_0) < \infty$. Using the method of successive approximations, one shows that, as $t \to +\infty$ and $x \to -\infty$ such that $\lambda_0 > M$, $\mu^c(\cdot)$ can be expanded in the following Neumann-type series (see also Part II of [23], and [34]):

$$\mathcal{A}_0(\varsigma)|_{\varsigma(\widetilde{w})} = \mu^c(\widetilde{w})|_{L'_s\setminus\{0\}} = \mu^c_{00}(\widetilde{w};\lambda_0) + \frac{\mu^c_{10}(\widetilde{w};\lambda_0) + \mu^c_{11}(\widetilde{w};\lambda_0)\ln t}{\sqrt{t}} + \mathcal{O}\left(\frac{\mu^c_{22}(\widetilde{w};\lambda_0)(\ln t)^2}{t}\right),$$

where $\mu^c_{00}(\widetilde{w};\lambda_0) := (\underline{\mathbf{Id}} - C^{\natural}_{00}(\widetilde{w};\lambda_0))^{-1}\mathrm{I}$, $||(\underline{\mathbf{Id}} - C^{\natural}_{00}(\cdot;\lambda_0))^{-1}\mathrm{I}||_{\mathcal{L}^2(L'_s\setminus\{0\};M_2(\mathbb{C}))} < \infty$, and $||\mu^c_{jk}(\cdot;\lambda_0)||_{\mathcal{L}^2(L'_s\setminus\{0\};M_2(\mathbb{C}))} < \infty$, $1 \le j \le 2$, $0 \le k \le j$: an explicit expression for $(\underline{\mathbf{Id}} - C^{\natural}_{00}(\widetilde{w};\lambda_0))^{-1}\mathrm{I}$ in terms of parabolic-cylinder functions was given in section 7 of [15] (see below). Making one more change of variable, $\varrho = \sqrt{2}\widetilde{w}\exp\{\frac{i\pi}{4}\}$, and recalling the definition of I_0, one shows that

$$I_0 - I_{1/2} = I_{0,a} + I_{0,b} + I_{0,c} + I_{0,d} + \mathcal{E}_r,$$

where

$$I_{1/2} := Y_a(\lambda,\lambda_0;t)\int_0^{\widehat{\alpha}} \mu^c_{00}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\varrho^{2i\nu}e^{-\varrho^2/2}\sigma_+ d\varrho,$$

$$I_{0,a} := Y_a(\lambda,\lambda_0;t)\sum_{\vec{l}_4\in\mathbb{L}_4}{}' \int_0^{\widehat{\alpha}} \mu^c_{00}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\prod_{k=1}^4\left(p_k\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\right)^{l_k}\varrho^{2i\nu}e^{-\varrho^2/2}\sigma_+ d\varrho,$$

$$I_{0,b} := Y_b(\lambda,\lambda_0;t)\sum_{\vec{l}_4\in\mathbb{L}_4} \int_0^{\widehat{\alpha}} \mu^c_{00}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\prod_{k=1}^4\left(p_k\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\right)^{l_k}\varrho^{2i\nu+1}e^{-\varrho^2/2}\sigma_+ d\varrho,$$

$$I_{0,c} := Y_c(\lambda,\lambda_0;t)\sum_{\vec{l}_4\in\mathbb{L}_4} \int_0^{\widehat{\alpha}} \left\{\mu^c_{10}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right) + \mu^c_{11}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\ln t\right\} \times \prod_{k=1}^4\left(p_k\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\right)^{l_k}\varrho^{2i\nu}e^{-\varrho^2/2}\sigma_+ d\varrho,$$

$$I_{0,d} := Y_d(\lambda,\lambda_0;t)\sum_{\vec{l}_4\in\mathbb{L}_4} \int_0^{\widehat{\alpha}} \mu^c_{00}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\prod_{k=1}^4\left(p_k\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\right)^{l_k}\varrho^{2i\nu+1}e^{-\varrho^2/2}\sigma_+ d\varrho,$$

$$\mathcal{E}_r := \mathcal{O}\left(\frac{y(\lambda_0;t)}{\lambda_0^3 t^{3/2}}\left\{\sum_{k=1}^2 \frac{C^{\sharp}_k(\lambda_0)}{(\lambda-\lambda_0)^k}\right\}\int_0^{\widehat{\alpha}} \mu^c_{00}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\varrho^{2i\nu+2}e^{-\varrho^2/2}\sigma_+ d\varrho\right)$$
$$+ \mathcal{O}\left(\frac{y(\lambda_0;t)}{\lambda_0^2 t^{3/2}}\left\{\sum_{k=1}^2 \frac{C^{\sharp}_{k+2}(\lambda_0)}{(\lambda-\lambda_0)^k}\right\}\int_0^{\widehat{\alpha}} \left\{\mu^c_{10}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right) + \mu^c_{11}\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\ln t\right\}\varrho^{2i\nu+1}e^{-\varrho^2/2}\sigma_+ d\varrho\right),$$

$\widehat{\alpha}:=(32\varepsilon^2\lambda_0^2 t)^{1/2}$, the prime on the summation in the expression for $I_{0,a}$ means that the term corresponding to $(l_1,l_2,l_3,l_4)=(0,0,0,0)$ is omitted from the sum,

$$Y_a(\lambda,\lambda_0;t)=\frac{y(\lambda_0;t)e^{-\frac{i\pi}{4}}\mathcal{R}(\lambda_0)}{(\lambda-\lambda_0)\sqrt{32\lambda_0^2 t}},\qquad Y_b(\lambda,\lambda_0;t)=\frac{iy(\lambda_0;t)\{3i\nu\mathcal{R}(\lambda_0)-\lambda_0\mathcal{R}'(\lambda_0)\}}{(\lambda-\lambda_0)(32\lambda_0^3 t)},$$

$$Y_c(\lambda,\lambda_0;t)=\frac{y(\lambda_0;t)e^{-\frac{i\pi}{4}}\mathcal{R}(\lambda_0)}{(\lambda-\lambda_0)\sqrt{32\lambda_0^2 t}},\qquad Y_d(\lambda,\lambda_0;t)=-\frac{iy(\lambda_0;t)\mathcal{R}(\lambda_0)}{(\lambda-\lambda_0)^2(32\lambda_0^2 t)},$$

$y(\lambda_0;t):=\frac{e^{\frac{\pi\nu}{2}}e^{2s(\lambda_0)}e^{4i\lambda_0^4 t}}{(2\pi i)\lambda_0^{4i\nu}2^{3i\nu}t^{i\nu}}$, and $C_i^\sharp(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$, $1\le i\le 4$. As will be shown below, $I_{1/2}$ gives rise to the leading-order ($\mathcal{O}(t^{-1/2})$) term: towards the proof of this statement, one proceeds by estimating the difference, $I_0-I_{1/2}$. Recall first the following inequality: $|\exp\{(\cdot)\}-1|\le|(\cdot)|\sup_{s\in[0,1]}|\exp\{s(\cdot)\}|$; hence, $|\exp\{\widetilde{\Delta}_i^\flat\}-1|\le|\widetilde{\Delta}_i^\flat|\sup_{s\in[0,1]}|\exp\{s\widetilde{\Delta}_i^\flat\}|$, where

$$\widetilde{\Delta}_i^\flat:=2\widetilde{\Delta}_i\left(\lambda_0+\frac{\varrho\exp\{-\frac{i\pi}{4}\}}{\sqrt{32\lambda_0^2 t}}\right),\qquad i\in\{1,2\}.$$

Since, as shown in [15], $||(\delta(\cdot))^{\pm1}||_{\mathcal{L}^\infty(\mathbb{C};\mathbb{C})}<\infty$, from the definitions of $\rho_\pm(\lambda_0)$, $\widehat{\rho}_\pm(\lambda_0)$, and $\widetilde{\Delta}_i^\flat$, $i\in\{1,2\}$, it follows that $\sup_{s\in[0,1]}|\exp\{s\widetilde{\Delta}_i^\flat\}|<\infty$; furthermore, using the Lipschitz property of $\ln\left(\frac{1-|r(\lambda)|^2}{1-|r(\lambda_0)|^2}\right)$, $|\lambda|<\lambda_0$, and the fact that $r(\lambda)\in\mathcal{S}(\widehat{\Gamma};\mathbb{C})$ and $||r||_{\mathcal{L}^\infty(\mathbb{R};\mathbb{C})}<1$, via an integration by parts argument, one deduces that

$$|\widetilde{\Delta}_1^\flat|\le\frac{K_1^\flat(\lambda_0)\varrho+K_2^\flat(\lambda_0)\varrho\ln\varrho+K_3^\flat(\lambda_0)\varrho\ln t}{\sqrt{\lambda_0^2 t}},\qquad |\widetilde{\Delta}_2^\flat|\le\frac{K_4^\flat(\lambda_0)\varrho}{\sqrt{\lambda_0^2 t}},$$

with $K_i^\flat(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{R}_{>0})$, $i\in\{1,4\}$. Similarly, one gets that

$$\left|\exp\left\{-\frac{i\varrho^3\exp(-\frac{3\pi i}{4})}{8\sqrt{2}\lambda_0^2\sqrt{t}}\right\}-1\right|\le\frac{\varrho^3}{8\sqrt{2}\lambda_0^2\sqrt{t}}\underbrace{\sup_{s\in[0,1]}\left|\exp\left\{\frac{-s(1-i)\varrho^3}{16\lambda_0^2\sqrt{t}}\right\}\right|}_{<\infty}:=\frac{\widetilde{K}_1(\lambda_0)\varrho^3}{\sqrt{\lambda_0^2 t}},$$

$$\left|\exp\left\{\frac{i\varrho^4}{4^4\lambda_0^4 t}\right\}-1\right|\le\frac{\varrho^4}{4^4\lambda_0^4 t}\underbrace{\sup_{s\in[0,1]}\left|\exp\left\{\frac{is\varrho^4}{4^4\lambda_0^4 t}\right\}\right|}_{<\infty}:=\frac{\widetilde{K}_2(\lambda_0)\varrho^4}{\lambda_0^2 t},$$

with $\widetilde{K}_i(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{R}_{>0})$, $i\in\{1,2\}$. Although the expression for the difference, $I_0-I_{1/2}$, contains many terms, estimations for the respective terms are analogous. Consider, say, the bound for the term corresponding to $(l_1,l_2,l_3,l_4)=(1,0,0,0)$ in $I_{0,a}$, which is denoted by $I_{0,a}^1$:

$$I_{0,a}^1:=Y_a(\lambda,\lambda_0;t)\int_0^{\widehat{\alpha}}\mu_{00}^c\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)p_1\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\varrho^{2i\nu}e^{-\varrho^2/2}\sigma_+ d\varrho.$$

Using the fact that $0<\nu\le\nu_{\max}:=-\frac{1}{2\pi}\ln(1-\sup_{\lambda\in\mathbb{R}}|r(\lambda)|^2)<\infty$, and recalling the definitions of $s(\lambda_0)$ and $\mathcal{R}(\lambda_0)$, one gets that, for $\lambda\in\mathbb{C}\setminus\mathcal{N}(\lambda_0;\epsilon_0)$, $|Y_a(\lambda,\lambda_0;t)|\le$

$\frac{\exp\{\frac{\pi\nu_{\max}}{2}\}|r(\lambda_0)|}{2\pi|\lambda-\lambda_0|\sqrt{32\lambda_0^2 t}}$; hence, letting the upper limit of integration tend to $+\infty$ (for brevity, the following notation is used: for *matrices* A and B, the inequality $|A|\leq|B|$ means that $|A_{ij}|\leq|B_{ij}|\ \forall\, i,j$),

$$|I_{0,a}^1|\leq\frac{\exp\{\frac{\pi\nu_{\max}}{2}\}|r(\lambda_0)|}{2\pi|\lambda-\lambda_0|\sqrt{32\lambda_0^2 t}}\int_0^\infty\left|\mu_{00}^c\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\right|\left|p_1\left(\frac{\varrho e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\right|e^{-\varrho^2/2}\sigma_+ d\varrho.$$

In [15], it was shown that

$$\mathcal{U}_{00}^c:=\left\|\mu_{00}^c\left(\frac{(\cdot)e^{-\frac{i\pi}{4}}}{\sqrt{2}};\lambda_0\right)\right\|_{\mathcal{L}^2(\mathbb{R}_{\geq 0};M_2(\mathbb{C}))}<\infty;$$

hence, from this estimate and the Cauchy–Schwarz inequality for integrals,

$$|I_{0,a}^1|\leq\frac{\exp\{\frac{\pi\nu_{\max}}{2}\}|r(\lambda_0)|\,\mathcal{U}_{00}^c\sigma_+}{2\pi|\lambda-\lambda_0|\sqrt{32\lambda_0^2 t}}\left\|p_1\left(\frac{(\cdot)e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\exp\{-(\cdot)^2/2\}\right\|_{\mathcal{L}^2(\mathbb{R}_{\geq 0};\mathbb{C})}.$$

Recalling the estimate for $|\exp\{\widetilde{\Delta}_1^\flat\}-1|$, one shows that

$$\left\|p_1\left(\frac{(\cdot)e^{-\frac{i\pi}{4}}}{\sqrt{2}}\right)\exp\{-(\cdot)^2/2\}\right\|_{\mathcal{L}^2(\mathbb{R}_{\geq 0};\mathbb{C})}\leq\frac{\widetilde{K}^\flat(\lambda_0)\ln t}{\sqrt{\lambda_0^2 t}},$$

where $\widetilde{K}^\flat(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{R}_{>0})$; hence, uniformly for $\lambda\in\mathbb{C}\backslash\mathcal{N}(\lambda_0;\epsilon_0)$,

$$|I_{0,a}^1|\leq\frac{\widetilde{K}_1^\sharp(\lambda_0)\ln t}{|\lambda-\lambda_0|\lambda_0^2 t},$$

with $\widetilde{K}_1^\sharp(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0}))$. Similarly, recalling the estimates for

$$\left|\exp\left\{-\frac{i\varrho^3\exp(-\frac{3\pi i}{4})}{8\sqrt{2}\lambda_0^2\sqrt{t}}\right\}-1\right|,\qquad\left|\exp\left\{\frac{i\varrho^4}{4^4\lambda_0^4 t}\right\}-1\right|,\quad\text{and}\quad|\exp\{\widetilde{\Delta}_2^\flat\}-1|,$$

and using the triangle inequality for $\mathcal{L}^2$-norms, one shows that the remaining terms for $I_{0,a}$ are of the type $\mathcal{O}(t^{-\frac{m}{2}})$ and $\mathcal{O}(t^{-\frac{n}{2}}\ln t)$, $2\leq m\leq 5$, $3\leq n\leq 6$. Estimating the remaining terms of $I_0-I_{1/2}$ analogously, one shows that, as $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0>M$, uniformly for $\lambda\in\mathbb{C}\backslash\mathcal{N}(\lambda_0;\epsilon_0)$,

$$|I_0-I_{1/2}|\leq\frac{\widetilde{K}_2^\sharp(\lambda;\lambda_0)\ln t}{\lambda_0^2 t},$$

where $||\widetilde{K}_2^\sharp(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C}\backslash\mathcal{N}(\lambda_0;\epsilon_0);M_2(\mathbb{R}_{>0}))}<\infty$, $\widetilde{K}_2^\sharp(\lambda;\cdot)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0}))$, and, as $\lambda\to\infty$, $\widetilde{K}_2^\sharp(\lambda;\lambda_0)\sim\mathcal{O}(\widetilde{k}_2^\sharp(\lambda_0)|\lambda|^{-1})$, with $\widetilde{k}_2^\sharp(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0}))$.

Repeating the whole of the above analysis mutatis mutandis for each term on the right-hand side of the original integral expression for $\chi^c(\lambda)$ which appears at the very beginning of the proof, one shows that, as $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0>M$, uniformly for $\lambda\in\mathbb{C}\backslash\cup_{\aleph\in\{0,\pm\lambda_0\}}\mathcal{N}(\aleph;\epsilon_0)$,

$$|\chi^c(\lambda)-\chi_{1/2}^c(\lambda)|\leq\frac{(h_1^+(\lambda;\lambda_0)+h_2^+(\lambda;\lambda_0))\ln t}{\lambda_0^2 t},$$

where $h_1^+(\lambda;\lambda_0):=\sum_{l'\in\{0,i0\}}e_{l'}^+(\lambda;\lambda_0)|r(l')|$, $h_2^+(\lambda;\lambda_0):=\sum_{l\in\{\pm\lambda_0\}}e_l^+(\lambda;\lambda_0)$, and the functions $e_{l'}^+(\lambda;\lambda_0)$ and $e_l^+(\lambda;\lambda_0)$ have the following property as $\lambda\to\infty$:

$$h_1^+(\lambda;\lambda_0)+h_2^+(\lambda;\lambda_0)\sim\mathcal{O}\left(\left\{\sum_{l'\in\{0,i0\}}e_{l'}^\sharp(\lambda_0)|r(l')|+\sum_{l\in\{\pm\lambda_0\}}e_l^\sharp(\lambda_0)\right\}|\lambda|^{-1}\right),$$
$$e_{l'}^\sharp(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0})),$$

$l'\in\{0,i0\}$, and $e_l^\sharp(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0}))$, $l\in\{\pm\lambda_0\}$; moreover,

$$||h_1^+(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C}\setminus\mathcal{N}(0;\epsilon_0);M_2(\mathbb{R}_{>0}))}<\infty,$$
$$||h_2^+(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C}\setminus\cup_{\aleph\in\{\pm\lambda_0\}}\mathcal{N}(\aleph;\epsilon_0);M_2(\mathbb{R}_{>0}))}<\infty, h_1^+(\lambda;\cdot)\in\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0})),$$
$$h_2^+(\lambda;\cdot)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{R}_{>0})),$$

and $\chi_{1/2}^c(\lambda)$ represents the sum over all $I_{1/2}$-like terms in which the upper limits of integration tend to $+\infty$. One can write $\chi_{1/2}^c(\lambda)$ in the following form,

$$\chi_{1/2}^c(\lambda)=\mathrm{I}+\frac{(\widetilde{\Lambda}^0)^{\mathrm{ad}(\sigma_3)}}{\sqrt{16\lambda_0^2t}}\mathcal{X}^{\Sigma_r}(\lambda)+\frac{(\widehat{\Lambda}_{\mathcal{C}}^0)^{\mathrm{ad}(\sigma_3)}}{\sqrt{8\lambda_0^2t}}\mathcal{X}^{\Sigma_{\mathcal{C}}}(\lambda),$$

where

$$\widetilde{\Lambda}^0 = \frac{\exp\{2i\lambda_0^4t\}}{(16\lambda_0^4t)^{\frac{i\nu}{2}}}\exp\left\{\sum_{l\in\{\pm\}}(\rho_l(\lambda_0)+\widehat{\rho}_l(\lambda_0))\right\},$$
$$\rho_\pm(\lambda_0) = \frac{1}{2\pi i}\int_0^{\pm\lambda_0}\ln\left(\frac{1-|r(\varsigma)|^2}{1-|r(\lambda_0)|^2}\right)\frac{d\varsigma}{(\varsigma-\lambda_0)},\qquad \widehat{\rho}_\pm(\lambda_0)=\int_{\pm i\infty}^{i0}\frac{\ln(1-r(\varsigma)\overline{r(\overline{\varsigma})})}{(\varsigma-\lambda_0)}\frac{d\varsigma}{2\pi i},$$
$$\mathcal{X}_{11}^{\Sigma_r}(\lambda) = \frac{\overline{r_{\mathcal{B}}(\lambda_0)}}{2^{2i\nu}}\int_0^{\varepsilon_1}\frac{\mathcal{X}_{+,12}^{\Sigma_{B,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}-\frac{\overline{r_{\mathcal{B}}(\lambda_0)}e^{2\pi\nu}}{2^{2i\nu}}\int_0^{\varepsilon_2}\frac{\mathcal{X}_{-,12}^{\Sigma_{B,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}$$
$$+\frac{\overline{r_{\mathcal{B}}(\lambda_0)}}{2^{2i\nu}}\int_0^{\varepsilon_1}\frac{\mathcal{X}_{+,12}^{\Sigma_{A,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i}-\frac{\overline{r_{\mathcal{B}}(\lambda_0)}e^{2\pi\nu}}{2^{2i\nu}}\int_0^{\varepsilon_2}\frac{\mathcal{X}_{-,12}^{\Sigma_{A,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i},$$
$$\mathcal{X}_{12}^{\Sigma_r}(\lambda) = -\frac{r_{\mathcal{B}}(\lambda_0)}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_1}}\frac{\mathcal{X}_{-,11}^{\Sigma_{B,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}+\frac{r_{\mathcal{B}}(\lambda_0)e^{2\pi\nu}}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_2}}\frac{\mathcal{X}_{+,11}^{\Sigma_{B,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}$$
$$-\frac{r_{\mathcal{B}}(\lambda_0)}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_1}}\frac{\mathcal{X}_{-,11}^{\Sigma_{A,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i}+\frac{r_{\mathcal{B}}(\lambda_0)e^{2\pi\nu}}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_2}}\frac{\mathcal{X}_{+,11}^{\Sigma_{A,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i},$$
$$\mathcal{X}_{21}^{\Sigma_r}(\lambda) = \frac{\overline{r_{\mathcal{B}}(\lambda_0)}}{2^{2i\nu}}\int_0^{\varepsilon_1}\frac{\mathcal{X}_{+,22}^{\Sigma_{B,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}-\frac{\overline{r_{\mathcal{B}}(\lambda_0)}e^{2\pi\nu}}{2^{2i\nu}}\int_0^{\varepsilon_2}\frac{\mathcal{X}_{-,22}^{\Sigma_{B,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}$$
$$+\frac{\overline{r_{\mathcal{B}}(\lambda_0)}}{2^{2i\nu}}\int_0^{\varepsilon_1}\frac{\mathcal{X}_{+,22}^{\Sigma_{A,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i}-\frac{\overline{r_{\mathcal{B}}(\lambda_0)}e^{2\pi\nu}}{2^{2i\nu}}\int_0^{\varepsilon_2}\frac{\mathcal{X}_{-,22}^{\Sigma_{A,r}}(\varsigma)\varsigma^{-2i\nu}e^{i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i},$$
$$\mathcal{X}_{22}^{\Sigma_r}(\lambda) = -\frac{r_{\mathcal{B}}(\lambda_0)}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_1}}\frac{\mathcal{X}_{-,21}^{\Sigma_{B,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}+\frac{r_{\mathcal{B}}(\lambda_0)e^{2\pi\nu}}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_2}}\frac{\mathcal{X}_{+,21}^{\Sigma_{B,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda-\lambda_0)}\frac{d\varsigma}{2\pi i}$$
$$-\frac{r_{\mathcal{B}}(\lambda_0)}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_1}}\frac{\mathcal{X}_{-,21}^{\Sigma_{A,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i}+\frac{r_{\mathcal{B}}(\lambda_0)e^{2\pi\nu}}{2^{-2i\nu}}\int_0^{\overline{\varepsilon_2}}\frac{\mathcal{X}_{+,21}^{\Sigma_{A,r}}(\varsigma)\varsigma^{2i\nu}e^{-i\varsigma^2}}{(\lambda+\lambda_0)}\frac{d\varsigma}{2\pi i},$$
$$r_{\mathcal{B}}(\lambda_0):=r(\lambda_0)\prod_{l=n+1}^{N}\left(\frac{(\lambda_0-\overline{\lambda_l})(\lambda_0+\overline{\lambda_l})}{(\lambda_0-\lambda_l)(\lambda_0+\lambda_l)}\right)^2,$$

$$\mathcal{X}^{\Sigma_{B,r}}_{-,11}(\varsigma)=\frac{\varsigma^{-i\nu}e^{\frac{i\varsigma^2}{2}}}{2^{\frac{i\nu}{2}}e^{-\frac{\pi\nu}{4}}}D_{i\nu}(\sqrt{2}\varsigma e^{\frac{i\pi}{4}}),\qquad \mathcal{X}^{\Sigma_{B,r}}_{+,11}(\varsigma)=\frac{\varsigma^{-i\nu}e^{\frac{i\varsigma^2}{2}}}{2^{\frac{i\nu}{2}}e^{\frac{3\pi\nu}{4}}}D_{i\nu}(\sqrt{2}\varsigma e^{-\frac{3\pi i}{4}}),$$

$$\mathcal{X}^{\Sigma_{B,r}}_{+,12}(\varsigma)=\frac{2^{\frac{i\nu}{2}}\varsigma^{i\nu}e^{-\frac{i\varsigma^2}{2}}e^{\frac{\pi\nu}{4}}}{\beta_{21}}\{\partial_\varsigma D_{-i\nu}(\sqrt{2}\varsigma e^{-\frac{i\pi}{4}})-i\varsigma D_{-i\nu}(\sqrt{2}\varsigma e^{-\frac{i\pi}{4}})\},$$

$$\mathcal{X}^{\Sigma_{B,r}}_{-,12}(\varsigma)=\frac{2^{\frac{i\nu}{2}}\varsigma^{i\nu}e^{-\frac{i\varsigma^2}{2}}e^{-\frac{3\pi\nu}{4}}}{\beta_{21}}\{\partial_\varsigma D_{-i\nu}(\sqrt{2}\varsigma e^{\frac{3\pi i}{4}})-i\varsigma D_{-i\nu}(\sqrt{2}\varsigma e^{\frac{3\pi i}{4}})\},$$

$$\mathcal{X}^{\Sigma_{B,r}}_{-,21}(\varsigma)=\frac{2^{-\frac{i\nu}{2}}\varsigma^{-i\nu}e^{\frac{i\varsigma^2}{2}}e^{\frac{\pi\nu}{4}}}{\beta_{12}}\{\partial_\varsigma D_{i\nu}(\sqrt{2}\varsigma e^{\frac{i\pi}{4}})+i\varsigma D_{i\nu}(\sqrt{2}\varsigma e^{\frac{i\pi}{4}})\},$$

$$\mathcal{X}^{\Sigma_{B,r}}_{+,21}(\varsigma)=\frac{2^{-\frac{i\nu}{2}}\varsigma^{-i\nu}e^{\frac{i\varsigma^2}{2}}e^{-\frac{3\pi\nu}{4}}}{\beta_{12}}\{\partial_\varsigma D_{i\nu}(\sqrt{2}\varsigma e^{-\frac{3\pi i}{4}})+i\varsigma D_{i\nu}(\sqrt{2}\varsigma e^{-\frac{3\pi i}{4}})\},$$

$$\mathcal{X}^{\Sigma_{B,r}}_{+,22}(\varsigma)=\frac{\varsigma^{i\nu}e^{-\frac{i\varsigma^2}{2}}}{2^{-\frac{i\nu}{2}}e^{-\frac{\pi\nu}{4}}}D_{-i\nu}(\sqrt{2}\varsigma e^{-\frac{i\pi}{4}}),\qquad \mathcal{X}^{\Sigma_{B,r}}_{-,22}(\varsigma)=\frac{\varsigma^{i\nu}e^{-\frac{i\varsigma^2}{2}}}{2^{-\frac{i\nu}{2}}e^{\frac{3\pi\nu}{4}}}D_{-i\nu}(\sqrt{2}\varsigma e^{\frac{3\pi i}{4}}),$$

$$\mathcal{X}^{\Sigma_{A,r}}_{\pm,11}(\varsigma)=\mathcal{X}^{\Sigma_{B,r}}_{\pm,11}(\varsigma),\qquad \mathcal{X}^{\Sigma_{A,r}}_{\pm,12}(\varsigma)=-\mathcal{X}^{\Sigma_{B,r}}_{\pm,12}(\varsigma),$$

$$\mathcal{X}^{\Sigma_{A,r}}_{\pm,21}(\varsigma)=-\mathcal{X}^{\Sigma_{B,r}}_{\pm,21}(\varsigma),\qquad \mathcal{X}^{\Sigma_{A,r}}_{\pm,22}(\varsigma)=\mathcal{X}^{\Sigma_{B,r}}_{\pm,22}(\varsigma),$$

$$\beta_{12}=-\frac{2^{1+i\nu}\sqrt{\pi}e^{-\frac{\pi\nu}{2}}e^{\frac{i\pi}{4}}}{\overline{r_{\mathcal{B}}(\lambda_0)}\,\Gamma(-i\nu)},\qquad \beta_{21}=\overline{\beta_{12}},$$

the integrals are evaluated along the rays $(0,\varepsilon_k)$ (and their complex conjugates), $\varepsilon_1:=\infty\exp\{\frac{i\pi}{4}\}$, $\varepsilon_2:=\infty\exp\{-\frac{3\pi i}{4}\}$, $D_{\pm i\nu}(\cdot)$ is the parabolic-cylinder function [32], and

$$\widehat{\Lambda}^0_{\mathcal{C}}\;=\;\exp\Big\{\sum_{l\in\{\pm\}}(\rho^{\mathcal{C}}_l(0)+\widehat{\rho}^{\mathcal{C}}_l(0))\Big\},$$

$$\rho^{\mathcal{C}}_{\pm}(0)=-\frac{1}{2\pi i}\int_0^{\pm\lambda_0}\ln|\varsigma|\,d\ln(1-|r(\varsigma)|^2),\qquad \widehat{\rho}^{\mathcal{C}}_{\pm}(0)=\int_{\pm\infty}^0\frac{\ln(1+|r(i\varsigma)|^2)}{\varsigma}\,\frac{d\varsigma}{2\pi i},$$

$$\mathcal{X}^{\Sigma_{\mathcal{C}}}_{11}(\lambda)\;=\;\widehat{\mathcal{R}}^{(-)}_{\mathcal{C}}(0)\int_0^{\overline{\varepsilon_1}}\frac{\mathcal{X}^{\Sigma^2_{\mathcal{C},r}}_{-,12}(\varsigma)e^{-i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i}-\widehat{\mathcal{R}}^{(-)}_{\mathcal{C}}(0)\int_0^{\overline{\varepsilon_2}}\frac{\mathcal{X}^{\Sigma^4_{\mathcal{C},r}}_{-,12}(\varsigma)e^{-i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i},$$

$$\mathcal{X}^{\Sigma_{\mathcal{C}}}_{12}(\lambda)\;=\;-\widehat{\mathcal{R}}^{(+)}_{\mathcal{C}}(0)\int_0^{\varepsilon_1}\frac{\mathcal{X}^{\Sigma^1_{\mathcal{C},r}}_{-,11}(\varsigma)e^{i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i}+\widehat{\mathcal{R}}^{(+)}_{\mathcal{C}}(0)\int_0^{\varepsilon_2}\frac{\mathcal{X}^{\Sigma^3_{\mathcal{C},r}}_{-,11}(\varsigma)e^{i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i},$$

$$\mathcal{X}^{\Sigma_{\mathcal{C}}}_{21}(\lambda)\;=\;\widehat{\mathcal{R}}^{(-)}_{\mathcal{C}}(0)\int_0^{\overline{\varepsilon_1}}\frac{\mathcal{X}^{\Sigma^2_{\mathcal{C},r}}_{-,22}(\varsigma)e^{-i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i}-\widehat{\mathcal{R}}^{(-)}_{\mathcal{C}}(0)\int_0^{\overline{\varepsilon_2}}\frac{\mathcal{X}^{\Sigma^4_{\mathcal{C},r}}_{-,22}(\varsigma)e^{-i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i},$$

$$\mathcal{X}^{\Sigma_{\mathcal{C}}}_{22}(\lambda)\;=\;-\widehat{\mathcal{R}}^{(+)}_{\mathcal{C}}(0)\int_0^{\varepsilon_1}\frac{\mathcal{X}^{\Sigma^1_{\mathcal{C},r}}_{-,21}(\varsigma)e^{i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i}+\widehat{\mathcal{R}}^{(+)}_{\mathcal{C}}(0)\int_0^{\varepsilon_2}\frac{\mathcal{X}^{\Sigma^3_{\mathcal{C},r}}_{-,21}(\varsigma)e^{i\varsigma^2}}{\lambda}\,\frac{d\varsigma}{2\pi i},$$

$$\widehat{\mathcal{R}}^{(+)}_{\mathcal{C}}(0)\;=\;(-1)^{2i\nu}\lambda_0^{4i\nu}\left\{\frac{r_{\mathcal{C}}(0)}{(1-|r(0)|^2)}-\frac{r_{\mathcal{C}}(i0)}{(1+|r(i0)|^2)}\right\},\qquad \widehat{\mathcal{R}}^{(-)}_{\mathcal{C}}(0)=\overline{\widehat{\mathcal{R}}^{(+)}_{\mathcal{C}}(0)},$$

$$r_{\mathcal{C}}(0):=r(0)\exp\left\{4i\sum_{l=n+1}^{N}\gamma_l\right\},\qquad r_{\mathcal{C}}(i0):=r(i0)\exp\left\{4i\sum_{l=n+1}^{N}\gamma_l\right\},$$

where $r(0):=(r(\lambda)|_{\lambda\in\mathbb{R}})|_{\lambda=0}$ and $r(i0):=(r(\lambda)|_{\lambda\in i\mathbb{R}})|_{\lambda=0}$: the explicit expressions for $\mathcal{X}^{\Sigma^i_{\mathcal{C},r}}_{-,jk}(\varsigma)$, $i\in\{1,2,3,4\}$, $j,k\in\{1,2\}$, are not written down here since they will

not actually be needed. Since [15] $(r(\lambda)|_{\lambda\in\mathbb{R}})|_{\lambda=0}=(r(\lambda)|_{\lambda\in i\mathbb{R}})|_{\lambda=0}=0$, $\widehat{\mathcal{R}}_{\mathcal{C}}^{(+)}(0)=\widehat{\mathcal{R}}_{\mathcal{C}}^{(-)}(0)=h_1^+(\lambda;\lambda_0)=0$, hence, $\mathcal{X}_{ij}^{\Sigma_C}(\lambda)=0$, $i,j\in\{1,2\}$. To obtain the expression for $\mathcal{X}_{ij}^{\Sigma_r}(\lambda)$, $i,j\in\{1,2\}$, given above, use was made of the explicit representation for $\mu_{00}^c(\cdot;\lambda_0):=(\underline{\mathbf{Id}}-C_{00}^\natural(\cdot;\lambda_0))^{-1}\mathrm{I}$ on $L'\cup\overline{L'}$ (recall the definition of $I_{1/2}$) in terms of parabolic-cylinder functions given in section 7 of [15]. Now, substituting the expressions given above for $\mathcal{X}_{\pm,ij}^{\Sigma_{B,r}}(\varsigma)$ and $\mathcal{X}_{\pm,ij}^{\Sigma_{A,r}}(\varsigma)$, $i,j\in\{1,2\}$, into the corresponding integrals for $\mathcal{X}_{ij}^{\Sigma_r}(\lambda)$, $i,j\in\{1,2\}$, and using the following identities [32], $\partial_\varsigma D_a(\varsigma)=\frac{1}{2}(aD_{a-1}(\varsigma)-D_{a+1}(\varsigma))$, $\varsigma D_a(\varsigma)=D_{a+1}(\varsigma)+aD_{a-1}(\varsigma)$, and $|\Gamma(i\nu)|^2=\pi/(\nu\sinh\pi\nu)$, as well as the following integral [32]:

$$\int_0^\infty \exp\left(-\frac{x^2}{4}\right)x^{a-1}D_{-b}(x)dx=\frac{\sqrt{\pi}\exp\{-\frac{1}{2}(a+b)\ln 2\}\Gamma(a)}{\Gamma(\frac{1}{2}(a+b)+\frac{1}{2})},\qquad \Re(a)>0,$$

one obtains the result stated in the lemma. □

PROPOSITION 4.2. *As $\lambda\to\infty$, $\lambda\in\mathbb{C}\setminus(\widehat{\Gamma}\cup(\cup_{i=1}^n(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})))$, $\chi(\lambda)$ has the following asymptotic expansion:*

$$\begin{aligned}\chi(\lambda)=\mathrm{I}+\frac{1}{2\lambda}\Bigg(&\left\{\overline{Q^\chi(x,t)}+4\sum_{i=1}^n\left(\beta_i-\frac{\chi_{21}^c(\overline{\lambda_i})}{\chi_{11}^c(\overline{\lambda_i})}\widehat{\delta}_i\right)\right\}\sigma_-\\ &+\left\{Q^\chi(x,t)+4\sum_{i=1}^n\left(\omega_i-\frac{\chi_{12}^c(\lambda_i)}{\chi_{22}^c(\lambda_i)}\alpha_i\right)\right\}\sigma_+\Bigg)+\mathcal{O}(\lambda^{-2}),\end{aligned}\tag{94}$$

where $\lim_{\lambda\to\infty}(\chi^c(x,t;\lambda))_{12}:=Q^\chi(x,t)/2\lambda$, $\{\alpha_i,\omega_i\}_{i=1}^n$ satisfy the following nondegenerate system of $2n$ linear inhomogeneous algebraic equations:

$$\begin{bmatrix}\widehat{\mathcal{A}}^+ & \widehat{\mathcal{B}}^+\\ \widehat{\mathcal{C}}^+ & \widehat{\mathcal{D}}^+\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_n\\ \omega_1\\ \omega_2\\ \vdots\\ \omega_n\end{bmatrix}=\begin{bmatrix}g_1^+\chi_{12}^c(\lambda_1)\\ g_2^+\chi_{12}^c(\lambda_2)\\ \vdots\\ g_n^+\chi_{12}^c(\lambda_n)\\ \overline{g_1^+}\chi_{11}^c(\overline{\lambda_1})\\ \overline{g_2^+}\chi_{11}^c(\overline{\lambda_2})\\ \vdots\\ \overline{g_n^+}\chi_{11}^c(\overline{\lambda_n})\end{bmatrix},\tag{95}$$

where, for $i,j\in\{1,2,\dots,n\}$, the $n\times n$ matrix blocks $\widehat{\mathcal{A}}^+$, $\widehat{\mathcal{B}}^+$, $\widehat{\mathcal{C}}^+$, and $\widehat{\mathcal{D}}^+$ are defined as follows,

$$\widehat{\mathcal{A}}_{ij}^+:=\begin{cases}\frac{\lambda_i+g_i^+\chi_{12}^c(\lambda_i)\chi_{22}^c(\lambda_i)+\lambda_i g_i^+\mathrm{W}(\chi_{12}^c(\lambda_i),\chi_{22}^c(\lambda_i))}{\lambda_i\chi_{22}^c(\lambda_i)}, & i=j,\\ -\frac{2g_i^+(-\lambda_i\chi_{22}^c(\lambda_i)\chi_{12}^c(\lambda_j)+\lambda_j\chi_{22}^c(\lambda_j)\chi_{12}^c(\lambda_i))}{\chi_{22}^c(\lambda_j)(\lambda_i^2-\lambda_j^2)}, & i\neq j,\end{cases}$$

$$\widehat{\mathcal{B}}_{ij}^+:=\begin{cases}-\frac{2g_i^+(\lambda_i\chi_{22}^c(\lambda_i)\chi_{11}^c(\overline{\lambda_i})-\overline{\lambda_i}\chi_{21}^c(\overline{\lambda_i})\chi_{12}^c(\lambda_i))}{\chi_{11}^c(\overline{\lambda_i})(\lambda_i^2-\overline{\lambda_i}^2)}, & i=j,\\ -\frac{2g_i^+(\lambda_i\chi_{22}^c(\lambda_i)\chi_{11}^c(\overline{\lambda_j})-\overline{\lambda_j}\chi_{21}^c(\overline{\lambda_j})\chi_{12}^c(\lambda_i))}{\chi_{11}^c(\overline{\lambda_j})(\lambda_i^2-\overline{\lambda_j}^2)}, & i\neq j,\end{cases}$$

$$\widehat{\mathcal{C}}_{ij}^+:=\begin{cases}-\frac{2\overline{g_i^+}(-\overline{\lambda_i}\chi_{21}^c(\overline{\lambda_i})\chi_{12}^c(\lambda_i)+\lambda_i\chi_{22}^c(\lambda_i)\chi_{11}^c(\overline{\lambda_i}))}{\chi_{22}^c(\lambda_i)(\overline{\lambda_i}^2-\lambda_i^2)}, & i=j,\\ -\frac{2\overline{g_i^+}(-\overline{\lambda_i}\chi_{21}^c(\overline{\lambda_i})\chi_{12}^c(\lambda_j)+\lambda_j\chi_{22}^c(\lambda_j)\chi_{11}^c(\overline{\lambda_i}))}{\chi_{22}^c(\lambda_j)(\overline{\lambda_i}^2-\lambda_j^2)}, & i\neq j,\end{cases}$$

$$\widehat{\mathcal{D}}^+_{ij} := \begin{cases} \frac{\overline{\lambda_i}-\overline{g_i^+}\chi^c_{21}(\overline{\lambda_i})\chi^c_{11}(\overline{\lambda_i})+\overline{\lambda_i}\,\overline{g_i^+}\mathrm{W}(\chi^c_{21}(\overline{\lambda_i}),\chi^c_{11}(\overline{\lambda_i}))}{\overline{\lambda_i}\chi^c_{11}(\overline{\lambda_i})}, & i=j,\\ -\frac{2\overline{g_i^+}(\overline{\lambda_i}\chi^c_{21}(\overline{\lambda_i})\chi^c_{11}(\overline{\lambda_j})-\overline{\lambda_j}\chi^c_{21}(\overline{\lambda_j})\chi^c_{11}(\overline{\lambda_i}))}{\chi^c_{11}(\overline{\lambda_j})(\overline{\lambda_i}^2-\overline{\lambda_j}^2)}, & i\neq j, \end{cases}$$

$\{\beta_i,\widehat{\delta}_i\}_{i=1}^n$ satisfy the following nondegenerate system of $2n$ linear inhomogeneous algebraic equations,

$$\begin{bmatrix} \boxed{\widehat{\mathcal{E}}^+} & \boxed{\widehat{\mathcal{F}}^+} \\ \boxed{\widehat{\mathcal{G}}^+} & \boxed{\widehat{\mathcal{H}}^+} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \\ \widehat{\delta}_1 \\ \widehat{\delta}_2 \\ \vdots \\ \widehat{\delta}_n \end{bmatrix} = \begin{bmatrix} g_1^+\chi^c_{22}(\lambda_1) \\ g_2^+\chi^c_{22}(\lambda_2) \\ \vdots \\ g_n^+\chi^c_{22}(\lambda_n) \\ \overline{g_1^+}\chi^c_{21}(\overline{\lambda_1}) \\ \overline{g_2^+}\chi^c_{21}(\overline{\lambda_2}) \\ \vdots \\ \overline{g_n^+}\chi^c_{21}(\overline{\lambda_n}) \end{bmatrix}, \tag{96}$$

where, for $i,j\in\{1,2,\dots,n\}$, the $n\times n$ matrix blocks $\widehat{\mathcal{E}}^+$, $\widehat{\mathcal{F}}^+$, $\widehat{\mathcal{G}}^+$, and $\widehat{\mathcal{H}}^+$ are defined as follows,

$$\widehat{\mathcal{E}}^+_{ij} := \begin{cases} \frac{\lambda_i-g_i^+\chi^c_{12}(\lambda_i)\chi^c_{22}(\lambda_i)+\lambda_i g_i^+\mathrm{W}(\chi^c_{12}(\lambda_i),\chi^c_{22}(\lambda_i))}{\lambda_i\chi^c_{22}(\lambda_i)}, & i=j,\\ \frac{2g_i^+(\lambda_j\chi^c_{12}(\lambda_j)\chi^c_{22}(\lambda_i)-\lambda_i\chi^c_{12}(\lambda_i)\chi^c_{22}(\lambda_j))}{\chi^c_{22}(\lambda_j)(\lambda_i^2-\lambda_j^2)}, & i\neq j, \end{cases}$$

$$\widehat{\mathcal{F}}^+_{ij} := \begin{cases} \frac{2g_i^+(\lambda_i\chi^c_{12}(\lambda_i)\chi^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\chi^c_{11}(\overline{\lambda_i})\chi^c_{22}(\lambda_i))}{\chi^c_{11}(\overline{\lambda_i})(\lambda_i^2-\overline{\lambda_i}^2)}, & i=j,\\ \frac{2g_i^+(\lambda_i\chi^c_{12}(\lambda_i)\chi^c_{21}(\overline{\lambda_j})-\overline{\lambda_j}\chi^c_{11}(\overline{\lambda_j})\chi^c_{22}(\lambda_i))}{\chi^c_{11}(\overline{\lambda_j})(\lambda_i^2-\overline{\lambda_j}^2)}, & i\neq j, \end{cases}$$

$$\widehat{\mathcal{G}}^+_{ij} := \begin{cases} \frac{2\overline{g_i^+}(\lambda_i\chi^c_{12}(\lambda_i)\chi^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\chi^c_{11}(\overline{\lambda_i})\chi^c_{22}(\lambda_i))}{\chi^c_{22}(\lambda_i)(\overline{\lambda_i}^2-\lambda_i^2)}, & i=j,\\ \frac{2\overline{g_i^+}(\lambda_j\chi^c_{12}(\lambda_j)\chi^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\chi^c_{11}(\overline{\lambda_i})\chi^c_{22}(\lambda_j))}{\chi^c_{22}(\lambda_j)(\overline{\lambda_i}^2-\lambda_j^2)}, & i\neq j, \end{cases}$$

$$\widehat{\mathcal{H}}^+_{ij} := \begin{cases} \frac{\overline{\lambda_i}+\overline{g_i^+}\chi^c_{11}(\overline{\lambda_i})\chi^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\,\overline{g_i^+}\mathrm{W}(\chi^c_{11}(\overline{\lambda_i}),\chi^c_{21}(\overline{\lambda_i}))}{\overline{\lambda_i}\chi^c_{11}(\overline{\lambda_i})}, & i=j,\\ \frac{2\overline{g_i^+}(\overline{\lambda_i}\chi^c_{11}(\overline{\lambda_i})\chi^c_{21}(\overline{\lambda_j})-\overline{\lambda_j}\chi^c_{11}(\overline{\lambda_j})\chi^c_{21}(\overline{\lambda_i}))}{\chi^c_{11}(\overline{\lambda_j})(\overline{\lambda_i}^2-\overline{\lambda_j}^2)}, & i\neq j, \end{cases}$$

with

$$g_j^+ := C_j e^{2i\lambda_j^2x+4i\lambda_j^4t}(\delta^+(\lambda_j;\lambda_0))^{-2}\prod_{l=n+1}^{N}\left(\frac{(\lambda_j-\lambda_l)(\lambda_j+\lambda_l)}{(\lambda_j-\overline{\lambda_l})(\lambda_j+\overline{\lambda_l})}\right)^2, \qquad 1\le j\le n,$$

$\delta^+(\lambda_k;\lambda_0)$, $k\in\{1,2,\dots,n\}$, given in Theorem 2.1, (18), *and $\mathrm{W}(\chi^c_{ij}(z),\chi^c_{i'j'}(z))$ is the Wronskian of $\chi^c_{ij}(\lambda)$ and $\chi^c_{i'j'}(\lambda)$ $(i,j,i',j'\in\{1,2\})$ evaluated at z: $\mathrm{W}(\chi^c_{ij}(z),\chi^c_{i'j'}(z)) := (\chi^c_{ij}(\lambda)\partial_\lambda\chi^c_{i'j'}(\lambda)-\chi^c_{i'j'}(\lambda)\partial_\lambda\chi^c_{ij}(\lambda))|_{\lambda=z}$.*

Proof. For $1\le i\le n$, set

$$\mathrm{res}(\chi(\lambda);\lambda_i) = \begin{pmatrix} \alpha_i & a_i \\ \beta_i & b_i \end{pmatrix}, \qquad \mathrm{res}(\chi(\lambda);\overline{\lambda_i}) = \begin{pmatrix} c_i & \omega_i \\ d_i & \widehat{\delta}_i \end{pmatrix}. \tag{97}$$

From (92), (93), (97), and the polar (residue) conditions in Lemma 3.3, one gets a system of linear algebraic equations for $\{\alpha_i,\beta_i,a_i,b_i,c_i,d_i,\omega_i,\widehat{\delta}_i\}_{i=1}^n$: from this system, one shows that, for $1\le i\le n$,

$$\left.\begin{aligned}
(\chi_{12}^c(\lambda_i)\alpha_i+\chi_{22}^c(\lambda_i)a_i)g_i^+=0 &\Rightarrow a_i=-\tfrac{\chi_{12}^c(\lambda_i)}{\chi_{22}^c(\lambda_i)}\alpha_i,\\
(\chi_{12}^c(\lambda_i)\beta_i+\chi_{22}^c(\lambda_i)b_i)g_i^+=0 &\Rightarrow b_i=-\tfrac{\chi_{12}^c(\lambda_i)}{\chi_{22}^c(\lambda_i)}\beta_i,\\
(\chi_{11}^c(\overline{\lambda_i})c_i+\chi_{21}^c(\overline{\lambda_i})\omega_i)\overline{g_i^+}=0 &\Rightarrow c_i=-\tfrac{\chi_{21}^c(\overline{\lambda_i})}{\chi_{11}^c(\overline{\lambda_i})}\omega_i,\\
(\chi_{11}^c(\overline{\lambda_i})d_i+\chi_{21}^c(\overline{\lambda_i})\widehat{\delta}_i)\overline{g_i^+}=0 &\Rightarrow d_i=-\tfrac{\chi_{21}^c(\overline{\lambda_i})}{\chi_{11}^c(\overline{\lambda_i})}\widehat{\delta}_i;
\end{aligned}\right\}. \tag{98}$$

Using (98), which show that the matrices $\{\mathrm{res}(\chi(\lambda);\lambda_i)\}_{i=1}^n$ and $\{\mathrm{res}(\chi(\lambda);\overline{\lambda_i})\}_{i=1}^n$, resp., are degenerate, one simplifies the resulting system of linear algebraic equations for $\{\alpha_i,\beta_i,a_i,b_i,c_i,d_i,\omega_i,\widehat{\delta}_i\}_{i=1}^n$ and obtains (95) and (96): the nondegeneracy of systems (95) and (96) is a consequence of the unique solvability of the original RH problem (Lemma 2.2). By substituting (97) into (92) and defining $Q^\chi(x,t)$ as in the proposition, one obtains, using (93), the result given in Lemma 4.1, and (98), the result given by (94). □

COROLLARY 4.1. *As $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0>M$ and $(x,t)\in\Omega_n$,*

$$Q(x,t)=\underset{-}{Q}{}_+^{\mathcal{C}}(x,t)+4i\sum_{j=1}^n\left(\omega_j-\frac{\chi_{12}^c(\lambda_j)}{\chi_{22}^c(\lambda_j)}\alpha_j\right)+\mathcal{O}(C(\lambda_0)\exp\{-abt\}),$$

where $\underset{-}{Q}{}_+^{\mathcal{C}}(x,t):=iQ^\chi(x,t)$, a and b are given in Lemma 3.3, and $C(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{C})$.

Proof. Since, from Lemma 3.2, $Q(x,t)=2i\lim_{\lambda\to\infty}(\lambda m^\sharp(x,t;\lambda))_{12}$, the result follows from Lemma 3.3 and Proposition 4.2. □

PROPOSITION 4.3. *As $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0>M$ and $(x,t)\in\Omega_n$,*

$$\begin{aligned}
Q(x,t)=Q_+^{\mathcal{C}}(x,t)&+4i\left(\omega_n-\frac{\chi_{12}^c(\lambda_n)}{\chi_{22}^c(\lambda_n)}\alpha_n\right)\\
&+\mathcal{O}\left(\frac{C_1(\lambda_0)\ln t}{t}\right)+\mathcal{O}(C_2(\lambda_0)e^{-a_0b_0t}),
\end{aligned} \tag{99}$$

where

$$\alpha_n=\frac{\widehat{a}_{12}\overline{g_n^+}\chi_{11}^c(\overline{\lambda_n})+\widehat{a}_{22}g_n^+\chi_{12}^c(\lambda_n)}{(\widehat{a}_{11}\widehat{a}_{22}-\widehat{a}_{12}\widehat{a}_{21})}, \tag{100}$$

$$\omega_n=\frac{\widehat{a}_{11}\overline{g_n^+}\chi_{11}^c(\overline{\lambda_n})+\widehat{a}_{21}g_n^+\chi_{12}^c(\lambda_n)}{(\widehat{a}_{11}\widehat{a}_{22}-\widehat{a}_{12}\widehat{a}_{21})}, \tag{101}$$

$$\widehat{a}_{11}:=\frac{\lambda_n+g_n^+\chi_{12}^c(\lambda_n)\chi_{22}^c(\lambda_n)+\lambda_ng_n^+\mathrm{W}(\chi_{12}^c(\lambda_n),\chi_{22}^c(\lambda_n))}{\lambda_n\chi_{22}^c(\lambda_n)}, \tag{102}$$

$$\widehat{a}_{12}:=\frac{2g_n^+(\lambda_n\chi_{22}^c(\lambda_n)\chi_{11}^c(\overline{\lambda_n})-\overline{\lambda_n}\chi_{21}^c(\overline{\lambda_n})\chi_{12}^c(\lambda_n))}{\chi_{11}^c(\overline{\lambda_n})(\lambda_n^2-\overline{\lambda_n}^2)}, \tag{103}$$

$$\widehat{a}_{21}:=\frac{2\overline{g_n^+}(\lambda_n\chi_{22}^c(\lambda_n)\chi_{11}^c(\overline{\lambda_n})-\overline{\lambda_n}\chi_{21}^c(\overline{\lambda_n})\chi_{12}^c(\lambda_n))}{\chi_{22}^c(\lambda_n)(\overline{\lambda_n}^2-\lambda_n^2)}, \tag{104}$$

$$\widehat{a}_{22}:=\frac{\overline{\lambda_n}-\overline{g_n^+}\chi_{21}^c(\overline{\lambda_n})\chi_{11}^c(\overline{\lambda_n})+\overline{\lambda_n}\,\overline{g_n^+}\mathrm{W}(\chi_{21}^c(\overline{\lambda_n}),\chi_{11}^c(\overline{\lambda_n}))}{\overline{\lambda_n}\chi_{11}^c(\overline{\lambda_n})}, \tag{105}$$

$Q_+^{\mathcal{C}}(x,t)$ *is given in Theorem* 2.1, (20)–(22) *and* (24), $a_0 := \min(a, 8\min\{\eta_l\}_{l=1}^{n-1})$ (>0), $b_0 := \min(b, \min\{|\xi_n - \xi_l|\}_{l=1}^{n-1})$, $C_1(\lambda_0) \in \mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$, *and* $C_2(\lambda_0) \in \mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{C})$.

Proof. Solving (95) as $t \to +\infty$ and $x \to -\infty$ such that $\lambda_0 > M$ and $(x,t) \in \Omega_n$ for $\{\alpha_i\}_{i=1}^n$ and $\{\omega_i\}_{i=1}^n$ via Cramer's rule, one shows that

$$\alpha_i,\ \omega_i \sim \mathcal{O}\left(\exp\left\{-a^\flat \min_{1\le l\le n-1} |\xi_n - \xi_l| t\right\}\right), \qquad 1 \le i \le n-1, \tag{106}$$

where $a^\flat := 8\min\{\eta_l\}_{l=1}^{n-1}$ (>0), and α_n and ω_n are given by (100) and (101): the result now follows from Corollary 4.1 and the estimates in (106). ∎

PROPOSITION 4.4. *As* $t \to +\infty$ *and* $x \to -\infty$ *such that* $\lambda_0 > M$ *and* $(x,t) \in \Omega_n$,

$$Q(x,t) = Q_{\text{as}}^+(x,t) + \mathcal{O}\left(\frac{C(\lambda_0)\ln t}{t}\right),$$

where $Q_{\text{as}}^+(x,t)$ *is given in Theorem* 2.1, (14)–(29), *and* $C(\lambda_0) \in \mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

Proof. For the proof substitute $\chi_{ij}^c(\cdot)$, $i,j \in \{1,2\}$, from Lemma 4.1 into (99)–(105) and neglect exponentially small terms. ∎

5. Asymptotic evaluation of $((\Psi^{-1}(x,t;0))_{11})^2$. In this section, the phase integral, $((\Psi^{-1}(x,t;0))_{11})^2$, which appears in the gauge transformation (Proposition 2.3, (9)) is evaluated asymptotically as $t \to +\infty$ $(x/t \sim \mathcal{O}(1))$.

LEMMA 5.1. *As* $t \to +\infty$ *and* $x \to -\infty$ *such that* $\lambda_0 > M$ *and* $(x,t) \in \Omega_n$,

$$\begin{aligned}((\Psi^{-1}(x,t;0))_{11})^2 = {}& \exp\{2\ln(\chi_{22}^c(0))\} \\ & \times \exp\left\{\frac{2i}{\pi}\left(\int_0^{\lambda_0} \frac{\ln(1-|r(\varrho)|^2)}{\varrho} d\varrho - \int_0^\infty \frac{\ln(1+|r(i\varrho)|^2)}{\varrho} d\varrho\right)\right\} \\ & \times \exp\left\{-4i\sum_{l=n+1}^N \gamma_l\right\} \exp\left\{2\ln\left(1 - \sum_{i=1}^n \left(\frac{2b_i}{\lambda_i} + \frac{2\widehat{\delta}_i}{\overline{\lambda_i}}\right)\right)\right\} \\ & + \mathcal{O}(C(\lambda_0)e^{-abt}),\end{aligned}$$

where $b_i = -\frac{\chi_{12}^c(\lambda_i)}{\chi_{22}^c(\lambda_i)}\beta_i$, $1 \le i \le n$, $\{\beta_i, \widehat{\delta}_i\}_{i=1}^n$ *satisfy* (96), a *and* b *are given in Lemma* 3.3, *and* $C(\lambda_0) \in \mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{C})$.

Proof. From Lemma 2.1, the proof of Lemma 3.1, Proposition 3.2, (85), Lemma 3.2, (86), and Lemma 3.3, one gets that

$$\Psi(x,t;0) = \chi(0)(\delta(0))^{\sigma_3} \prod_{l=n+1}^N (d_{l_+}(0))^{\sigma_3} + \mathcal{O}(C_1(\lambda_0)\exp\{-abt\}),$$

where $C_1(\lambda_0) \in \mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{C}))$. From Propositions 2.1–2.3, the parametrization for the discrete eigenvalues (section 2), Lemma 3.2 (equation (87)), Proposition 4.1 (equations (92) and (93)), the proof of Proposition 4.2 (equations (97) and (98)), and the σ_1 and σ_3 symmetry reductions for $\chi(\lambda)$, one shows that,

$$\Psi^{-1}(x,t;0) = (\widehat{h}(0))^{\sigma_3} \begin{pmatrix} 1 - \sum_{i=1}^n \left(\frac{2b_i}{\lambda_i} + \frac{2\widehat{\delta}_i}{\overline{\lambda_i}}\right) & 0 \\ 0 & 1 - \sum_{i=1}^n \left(\frac{2\alpha_i}{\lambda_i} + \frac{2c_i}{\overline{\lambda_i}}\right) \end{pmatrix} + \mathcal{O}(C_2(\lambda_0)e^{-abt}),$$

where $\widehat{h}(0) := \chi^c_{22}(0)(\delta(0))^{-1}\exp\{-2i\sum_{l=n+1}^{N}\gamma_l\}$, b_i, $1\le i\le n$, and $\{\beta_i,\widehat{\delta}_i\}_{i=1}^n$ are as given in the lemma, $c_i = -\frac{\chi^c_{21}(\overline{\lambda_i})}{\chi^c_{11}(\lambda_i)}\omega_i$, $1\le i\le n$, $\{\alpha_i,\omega_i\}_{i=1}^n$ are defined by system (95), and $C_2(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{C}))$: using the expression for $\delta^+(\lambda;\lambda_0)$ given in Proposition 3.1 (and Remark 3.1), one obtains the result stated in the lemma. ∎

In order to estimate $(\chi^c_{22}(0))^2$, the following proposition and lemma are necessary.

PROPOSITION 5.1. *Define* $Q^\sharp(x,t) := 2i\lim_{\lambda\to\infty}(\lambda\chi^c(x,t;\lambda))_{12}$. *Then*

$$(||Q^\sharp(\cdot,t)||_{\mathcal{L}^2(\mathbb{R};\mathbb{C})})^2 = \frac{2}{\pi}\left(\int_0^\infty \frac{\ln(1+|r(i\varrho)|^2)}{\varrho}d\varrho - \int_0^\infty \frac{\ln(1-|r(\varrho)|^2)}{\varrho}d\varrho\right),$$
$$(\chi^c_{22}(0))^2 = (\delta^+(0;\lambda_0))^2\exp\left\{i\int_{+\infty}^x |Q^\sharp(\varrho,t)|^2 d\varrho\right\}.$$

Proof. The proof follows from the definition of $\chi^c(\lambda)$ given in Proposition 4.1, Proposition 2.2, and Proposition 8.1 in [15]. ∎

LEMMA 5.2 (see [35]). *As* $t\to+\infty$ *and* $x\to-\infty$ *such that* $\lambda_0>M$,

$$Q^\sharp(x,t) = \frac{u^+_{1,1,0}(\lambda_0)e^{i(4\lambda_0^4 t-\nu(\lambda_0)\ln t)}}{\sqrt{t}} + \frac{u^+_{-1,2,0}(\lambda_0)}{t} + \mathcal{O}\left(\frac{C(\lambda_0)(\ln t)^2}{t^{3/2}}\right),$$

where

$$u^+_{1,1,0}(\lambda_0) = \sqrt{\frac{\nu(\lambda_0)}{2\lambda_0^2}}\exp\{i\theta^+(\lambda_0)\},$$
$$\theta^+(\lambda_0) = \phi^+(\lambda_0) - \frac{3\pi}{4} + \arg\Gamma(i\nu(\lambda_0)) + \arg r(\lambda_0) - 3\nu(\lambda_0)\ln 2 + 2\sum_{l=n+1}^{N}\arg\left(\frac{(\lambda_0-\overline{\lambda_l})(\lambda_0+\overline{\lambda_l})}{(\lambda_0-\lambda_l)(\lambda_0+\lambda_l)}\right),$$
$$u^+_{-1,2,0}(\lambda_0) = -\frac{i}{8\pi\lambda_0^2}\left(\left.\frac{d(r(\varrho)|_{\varrho\in\mathbb{R}})}{d\varrho}\right|_{\varrho=0} - \left.\frac{d(r(\varrho)|_{\varrho\in i\mathbb{R}})}{d\varrho}\right|_{\varrho=0}\right) \times\exp\left\{i\left(4\sum_{l=n+1}^{N}\gamma_l + 2\vartheta^+(\lambda_0)\right)\right\},$$
$$\vartheta^+(\lambda_0) = -\int_0^{\lambda_0}\frac{\ln(1-|r(\varrho)|^2)}{\varrho}\frac{d\varrho}{\pi} + \int_0^\infty \frac{\ln(1+|r(i\varrho)|^2)}{\varrho}\frac{d\varrho}{\pi},$$

$\phi^+(\cdot)$ *is given in Theorem* 2.1, (22), *and* $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

Comment to proof. Up to the leading ($\mathcal{O}(t^{-1})$) term, the asymptotic expansion was proved in [15]. The $\mathcal{O}(t^{-1})$ term constitutes the leading-order contribution from the first-order stationary phase point at $\lambda=0$: the complete proof of this asymptotic expansion can be found in [35].

PROPOSITION 5.2. *As* $t\to+\infty$ *and* $x\to-\infty$ *such that* $\lambda_0>M$,

$$(\chi^c_{22}(0))^2 = \exp\left\{i\left(\sqrt{\frac{2}{t}}\int_{\lambda_0}^\infty \frac{\sqrt{\nu(\mu)}}{\mu^2}\left(R_i^+(0)\cos(\kappa^+(\mu;t)) - R_r^+(0)\sin(\kappa^+(\mu;t))\right)\frac{d\mu}{\pi}\right)\right\} + \mathcal{O}\left(\frac{C(\lambda_0)(\ln t)^2}{\lambda_0^2 t}\right),$$

where

$$R_i^+(0) = \Im\{R^+(0)\}, \qquad R_r^+(0) = \Re\{R^+(0)\}$$

$$:= \left(\frac{d(r(\varrho)|_{\varrho\in\mathbb{R}})}{d\varrho}\bigg|_{\varrho=0} - \frac{d(r(\varrho)|_{\varrho\in i\mathbb{R}})}{d\varrho}\bigg|_{\varrho=0} \right) \exp\left\{ 4i \sum_{l=n+1}^{N} \gamma_l \right\},$$

$$\kappa^+(\lambda_0;t) := 4\lambda_0^4 t - \nu(\lambda_0)\ln t + \theta^+(\lambda_0) - 2\vartheta^+(\lambda_0),$$

and $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

Proof. Writing $\int_{+\infty}^{x}|Q^\sharp(\varrho,t)|^2 d\varrho = -(||Q^\sharp(\cdot,t)||_{\mathcal{L}^2(\mathbb{R};\mathbb{C})})^2 + \int_{-\infty}^{x}|Q^\sharp(\varrho,t)|^2 d\varrho$, using the expressions for $(||Q^\sharp(\cdot,t)||_{\mathcal{L}^2(\mathbb{R};\mathbb{C})})^2$ and $(\chi_{22}^c(0))^2$ given in Proposition 5.1, the asymptotic expansion for $Q^\sharp(x,t)$ given in Lemma 5.2, the following inequalities: $|\exp\{(\cdot)\}-1| \le |(\cdot)|\sup_{s\in[0,1]}|\exp\{s(\cdot)\}|$ and $0<\nu(\lambda_0)\le\nu_{\max} := -\frac{1}{2\pi}\ln(1-\sup_{\lambda\in\mathbb{R}}|r(\lambda)|^2)<\infty$, and the fact that $r(\lambda)\in\mathcal{S}(\widehat{\Gamma};\mathbb{C})$, one obtains the result stated in the proposition. ∎

LEMMA 5.3. *As* $t\to+\infty$ *and* $x\to-\infty$ *such that* $\lambda_0>M$ *and* $(x,t)\in\Omega_n$,

$$((\Psi^{-1}(x,t;0))_{11})^2 = \exp\{i\arg q_{\text{as}}^+(x,t)\} + \mathcal{O}\left(\frac{C(\lambda_0)(\ln t)^2}{t}\right),$$

where $\arg q_{\text{as}}^+(x,t)$ *is given in Theorem* 2.2, (48)–(51), *and* $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

Proof. According to Lemma 5.1, in order to evaluate $((\Psi^{-1}(x,t;0))_{11})^2$, estimates for $\exp\{2\ln(\chi_{22}^c(0))\}$ and $\{b_i,\widehat{\delta}_i\}_{i=1}^n$ are required: the estimation for $\exp\{2\ln(\chi_{22}^c(0))\}$ is given in Proposition 5.2; hence, it remains to estimate $\{b_i,\widehat{\delta}_i\}_{i=1}^n$. Solving system (96) as $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0>M$ and $(x,t)\in\Omega_n$ for $\{\beta_i\}_{i=1}^n$ and $\{\widehat{\delta}_i\}_{i=1}^n$ via Cramer's rule, one shows that β_i, $\widehat{\delta}_i\sim\mathcal{O}(\exp\{-a^\flat\min_{1\le i\le n-1}|\xi_n-\xi_i|t\})$, $1\le i\le n-1$, and

$$\beta_n = \frac{\beta_n^+}{(\widehat{\mathcal{E}}_{nn}^+\widehat{\mathcal{H}}_{nn}^+ - \widehat{\mathcal{F}}_{nn}^+\widehat{\mathcal{G}}_{nn}^+)}, \qquad \widehat{\delta}_n = \frac{\widehat{\delta}_n^+}{(\widehat{\mathcal{E}}_{nn}^+\widehat{\mathcal{H}}_{nn}^+ - \widehat{\mathcal{F}}_{nn}^+\widehat{\mathcal{G}}_{nn}^+)},$$

where

$$\begin{aligned}
\beta_n^+ :=\ & \frac{g_n^+\chi_{22}^c(\lambda_n)}{\chi_{11}^c(\overline{\lambda_n})} + \frac{|g_n^+|^2\chi_{21}^c(\overline{\lambda_n})\chi_{22}^c(\lambda_n)}{\overline{\lambda_n}} - \frac{|g_n^+|^2\chi_{22}^c(\lambda_n)W(\chi_{11}^c(\overline{\lambda_n}),\chi_{21}^c(\overline{\lambda_n}))}{\chi_{11}^c(\overline{\lambda_n})} \\
& + \frac{2\overline{\lambda_n}|g_n^+|^2\chi_{22}^c(\lambda_n)\chi_{21}^c(\overline{\lambda_n})}{(\lambda_n^2-\overline{\lambda_n}^2)}, \\
\widehat{\mathcal{E}}_{nn}^+\widehat{\mathcal{H}}_{nn}^+ - \widehat{\mathcal{F}}_{nn}^+\widehat{\mathcal{G}}_{nn}^+ :=\ & \frac{1}{\chi_{22}^c(\lambda_n)\chi_{11}^c(\overline{\lambda_n})} + \frac{\overline{g_n^+}W(\chi_{21}^c(\overline{\lambda_n}),\chi_{11}^c(\overline{\lambda_n}))}{\chi_{22}^c(\lambda_n)\chi_{11}^c(\overline{\lambda_n})} \\
& + \frac{g_n^+W(\chi_{12}^c(\lambda_n),\chi_{22}^c(\lambda_n))}{\chi_{11}^c(\overline{\lambda_n})\chi_{22}^c(\lambda_n)} + \frac{\overline{g_n^+}\chi_{21}^c(\overline{\lambda_n})}{\overline{\lambda_n}\chi_{22}^c(\lambda_n)} - \frac{g_n^+\chi_{12}^c(\lambda_n)}{\lambda_n\chi_{11}^c(\overline{\lambda_n})} \\
& + \frac{(2\overline{\lambda_n})^2|g_n^+|^2\chi_{22}^c(\lambda_n)\chi_{11}^c(\overline{\lambda_n})}{(\lambda_n^2-\overline{\lambda_n}^2)^2}, \widehat{\delta}_n^+ \\
:=\ & \frac{\overline{g_n^+}\chi_{21}^c(\overline{\lambda_n})}{\chi_{22}^c(\lambda_n)} - \frac{2\overline{\lambda_n}|g_n^+|^2\chi_{11}^c(\overline{\lambda_n})\chi_{22}^c(\lambda_n)}{(\lambda_n^2-\overline{\lambda_n}^2)}.
\end{aligned}$$

Substituting the expressions for $\chi^c_{ij}(\cdot)$, $i,j\in\{1,2\}$, given in Lemma 4.1 into the above equations for β_n^+, $\widehat{\delta}_n^+$, and $\widehat{\mathcal{E}}^+_{nn}\widehat{\mathcal{H}}^+_{nn}-\widehat{\mathcal{F}}^+_{nn}\widehat{\mathcal{G}}^+_{nn}$, and recalling that (see (98)) $b_i=-\frac{\chi^c_{12}(\lambda_i)}{\chi^c_{22}(\lambda_i)}\beta_i$, $1\le i\le n$, one obtains, as a result of Lemma 5.1, keeping only $\mathcal{O}(1)$ and $\mathcal{O}(t^{-1/2})$ terms, the result stated in the lemma. □

COROLLARY 5.1. *As $t\to+\infty$ and $x\to-\infty$ such that $\lambda_0>M$ and $(x,t)\in\Omega_n$,*

$$q(x,t)=Q^+_{\rm as}(x,t)\exp\{i\arg q^+_{\rm as}(x,t)\}+\mathcal{O}\bigg(\frac{C(\lambda_0)(\ln t)^2}{t}\bigg),$$

where $Q^+_{\rm as}(x,t)$ is given in Theorem 2.1, (14)–(29), *$\arg q^+_{\rm as}(x,t)$ is given in Theorem* 2.2, (48)–(51), *and $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.*

Proof. The proof is a consequence of Proposition 2.3 and Lemma 5.3. □

COROLLARY 5.2. *As $t\to+\infty$ and $x\to+\infty$ such that $\widehat{\lambda}_0:=\sqrt{\frac{1}{2}(\frac{x}{t}-\frac{1}{s})}>M$, $\frac{x}{t}>\frac{1}{s}$, $s\in\mathbb{R}_{>0}$, and $(x,t)\in\widetilde{\Omega}_n$,*

$$u(x,t)=v^+_{\rm as}(x,t)w^+_{\rm as}(x,t)+\mathcal{O}\bigg(\frac{C(\widehat{\lambda}_0)(\ln t)^2}{t}\bigg),$$

where $v^+_{\rm as}(x,t)$ and $w^+_{\rm as}(x,t)$ are given in Theorem 2.3, (58)–(70), *and $C(\widehat{\lambda}_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.*

Proof. The proof is a consequence of Proposition 2.4 and Corollary 5.1. □

6. Asymptotics as $t\to-\infty$. In this section, the asymptotic paradigm presented in sections 3–5 is reworked for the case when $t\to-\infty$: since the proofs of all obtained results are analogous, they will be omitted. This section is divided into three parts: (1) in subsection 6.1, extended and model RH problems are formulated as $t\to-\infty$; (2) in subsection 6.2, the model RH problem formulated in (1) above is solved asymptotically as $t\to-\infty$ for the Schwartz class of nonreflectionless generic potentials; and (3) in subsection 6.3, the phase integral, $((\Psi^{-1}(x,t;0))_{11})^2$, is evaluated asymptotically as $t\to-\infty$.

6.1. Extended and model RH problems.

PROPOSITION 6.1.1. *In the solitonless sector ($\mathcal{Z}_d\equiv\emptyset$), as $t\to-\infty$ and $x\to+\infty$ such that $\lambda_0>M$,*

$$m(x,t;\lambda)=\widetilde{\Delta}(\lambda)+\mathcal{O}\bigg(\frac{C(\lambda_0)}{\sqrt{-t}}\bigg),$$

where $\widetilde{\Delta}(\lambda):=(\delta^-(\lambda;\lambda_0))^{\sigma_3}$,

$$\delta^-(\lambda;\lambda_0)=((\lambda-\lambda_0)(\lambda+\lambda_0))^{-i\nu}\exp\bigg\{\sum_{l\in\{\pm\}}\widetilde{\rho}_l(\lambda)\bigg\},$$

$$\widetilde{\rho}_\pm(\lambda)=-\frac{1}{2\pi i}\int_{\pm\lambda_0}^{\pm\infty}\ln(\varsigma-\lambda)d\ln(1-|r(\varsigma)|^2),$$

$\nu:=\nu(\lambda_0)$ is given by (21), *$||(\delta^-(\cdot;\lambda_0))^{\pm1}||_{\mathcal{L}^\infty(\mathbb{C};\mathbb{C})}:=\sup_{\lambda\in\mathbb{C}}|(\delta^-(\lambda;\lambda_0))^{\pm1}|<\infty$, $\overline{(\delta^-(\pm\overline{\lambda};\lambda_0))^{-1}}=\delta^-(\lambda;\lambda_0)$, the principal branch of the logarithmic function is taken, $\ln(\mu-\lambda):=\ln|\mu-\lambda|+i\arg(\mu-\lambda)$, $\arg(\mu-\lambda)\in(-\pi,\pi)$, and $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{C}))$.*

Remark 6.1.1. Hereafter, all explicit x, t dependencies are suppressed, except where absolutely necessary, and $\delta^-(\lambda;\lambda_0) := \widetilde{\delta}(\lambda)$.

LEMMA 6.1.1. *There exists a unique solution* $\widetilde{m}^{\widetilde{\Delta}}(\lambda) := m(\lambda)(\widetilde{\Delta}(\lambda))^{-1}: \mathbb{C}\setminus(\mathcal{Z}_d \cup \widehat{\Gamma}) \to \mathrm{SL}(2,\mathbb{C})$ *of the following RH problem:*

1. $\widetilde{m}^{\widetilde{\Delta}}(\lambda)$ *is meromorphic for all* $\lambda \in \mathbb{C}\setminus\widehat{\Gamma}$,
2. $$\widetilde{m}^{\widetilde{\Delta}}_+(\lambda) = \widetilde{m}^{\widetilde{\Delta}}_-(\lambda)\widetilde{v}^{\widetilde{\Delta}}(\lambda), \quad \lambda \in \widehat{\Gamma},$$
where
$$\widetilde{v}^{\widetilde{\Delta}}(\lambda) = e^{-i\theta(\lambda)\mathrm{ad}(\sigma_3)} \begin{pmatrix} (1 - r(\lambda)\overline{r(\overline{\lambda})})\widetilde{\delta}_-(\lambda)(\widetilde{\delta}_+(\lambda))^{-1} & r(\lambda)\widetilde{\delta}_-(\lambda)\widetilde{\delta}_+(\lambda) \\ -\overline{r(\overline{\lambda})}(\widetilde{\delta}_-(\lambda))^{-1}(\widetilde{\delta}_+(\lambda))^{-1} & (\widetilde{\delta}_-(\lambda))^{-1}\widetilde{\delta}_+(\lambda) \end{pmatrix},$$
3. $\widetilde{m}^{\widetilde{\Delta}}(\lambda)$ *has simple poles at* $\{\pm\lambda_i, \pm\overline{\lambda_i}\}_{i=1}^N$ *with* $(1 \le i \le N)$
$$\begin{aligned} \mathrm{res}(\widetilde{m}^{\widetilde{\Delta}}(\lambda);\lambda_i) &= \lim_{\lambda\to\lambda_i} \widetilde{m}^{\widetilde{\Delta}}(\lambda) v_i (\widetilde{\delta}(\lambda_i))^{-2}\sigma_-, \\ \mathrm{res}(\widetilde{m}^{\widetilde{\Delta}}(\lambda);-\lambda_i) &= -\sigma_3 \mathrm{res}(\widetilde{m}^{\widetilde{\Delta}}(\lambda);\lambda_i)\sigma_3, \\ \mathrm{res}(\widetilde{m}^{\widetilde{\Delta}}(\lambda);\overline{\lambda_i}) &= \lim_{\lambda\to\overline{\lambda_i}} \widetilde{m}^{\widetilde{\Delta}}(\lambda) \overline{v_i} (\widetilde{\delta}(\overline{\lambda_i}))^{2}\sigma_+, \\ \mathrm{res}(\widetilde{m}^{\widetilde{\Delta}}(\lambda);-\overline{\lambda_i}) &= -\sigma_3 \mathrm{res}(\widetilde{m}^{\widetilde{\Delta}}(\lambda);\overline{\lambda_i})\sigma_3, \end{aligned}$$
4. *as* $\lambda \to \infty$, $\lambda \in \mathbb{C}\setminus(\mathcal{Z}_d \cup \widehat{\Gamma})$,
$$\widetilde{m}^{\widetilde{\Delta}}(\lambda) = \mathrm{I} + \mathcal{O}(\lambda^{-1});$$

moreover, $Q(x,t) = 2i\lim_{\lambda\to\infty}(\lambda \widetilde{m}^{\widetilde{\Delta}}(x,t;\lambda))_{12}$ *is equal to* $Q(x,t)$ *in Lemma* 2.2, (11).

PROPOSITION 6.1.2. *Introduce arbitrarily small, clockwise- and counterclockwise-oriented, mutually disjoint (and disjoint with respect to* $\widehat{\Gamma}$*) circles* $\widetilde{K}_j^{\pm}$ *and* $\widetilde{L}_j^{\pm}$, $1 \le j \le n-1$, *around the eigenvalues* $\{\pm\lambda_j\}_{j=1}^{n-1}$ *and* $\{\pm\overline{\lambda_j}\}_{j=1}^{n-1}$, *resp., and define*

$$\widetilde{m}^{\flat}(\lambda) := \begin{cases} \widetilde{m}^{\widetilde{\Delta}}(\lambda), \quad \lambda \in \mathbb{C}\Big\backslash \left(\widehat{\Gamma} \cup \left(\bigcup_{i=1}^{n-1} (\widetilde{K}_i^{\pm} \cup \widetilde{L}_i^{\pm})\right)\right), \\ \widetilde{m}^{\widetilde{\Delta}}(\lambda)\left(\mathrm{I} - \frac{v_i(\widetilde{\delta}(\pm\lambda_i))^{-2}}{(\lambda \mp \lambda_i)}\sigma_-\right), \quad \lambda \in \mathrm{int}\widetilde{K}_i^{\pm}, \quad 1 \le i \le n-1, \\ \widetilde{m}^{\widetilde{\Delta}}(\lambda)\left(\mathrm{I} + \frac{\overline{v_i}(\widetilde{\delta}(\pm\overline{\lambda_i}))^{2}}{(\lambda \mp \overline{\lambda_i})}\sigma_+\right), \quad \lambda \in \mathrm{int}\widetilde{L}_i^{\pm}, \quad 1 \le i \le n-1. \end{cases}$$

Then $\widetilde{m}^{\flat}(\lambda)$ *solves an RH problem on* $(\sigma_{\mathcal{L}}\setminus\cup_{i=1}^{n-1}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\cup_{i=1}^{n-1}(\widetilde{K}_i^{\pm}\cup\widetilde{L}_i^{\pm}))$ *with the same jumps as* $\widetilde{m}^{\widetilde{\Delta}}(\lambda)$ *on* $\widehat{\Gamma}$, $\widetilde{m}^{\flat}_+(\lambda) = \widetilde{m}^{\flat}_-(\lambda)\widetilde{v}^{\widetilde{\Delta}}(\lambda)$, *and*

$$\widetilde{m}^{\flat}_+(\lambda) = \begin{cases} \widetilde{m}^{\flat}_-(\lambda)\left(\mathrm{I} + \frac{v_i(\widetilde{\delta}(\pm\lambda_i))^{-2}}{(\lambda \mp \lambda_i)}\sigma_-\right), \quad \lambda \in \widetilde{K}_i^{\pm}, \quad 1 \le i \le n-1, \\ \widetilde{m}^{\flat}_-(\lambda)\left(\mathrm{I} + \frac{\overline{v_i}(\widetilde{\delta}(\pm\overline{\lambda_i}))^{2}}{(\lambda \mp \overline{\lambda_i})}\sigma_+\right), \quad \lambda \in \widetilde{L}_i^{\pm}, \quad 1 \le i \le n-1. \end{cases}$$

Remark 6.1.2. The superscripts $\pm$ on $\{\widetilde{K}_i^{\pm}\}_{i=1}^{n-1}$ and $\{\widetilde{L}_i^{\pm}\}_{i=1}^{n-1}$, which are related with $\{\pm\lambda_i\}_{i=1}^{n-1}$ and $\{\pm\overline{\lambda_i}\}_{i=1}^{n-1}$, resp., should *not* be confused with the subscripts $\pm$

appearing in the various RH problems in this and the next subsection, namely, $m_\pm(\lambda)$, $\widetilde{m}^{\widetilde{\Delta}}_\pm(\lambda)$, $\widetilde{m}^\flat_\pm(\lambda)$, $\widetilde{m}^\sharp_\pm(\lambda)$, $\widetilde{\chi}_\pm(\lambda)$, $\widetilde{E}_\pm(\lambda)$, and $\widetilde{\chi}^c_\pm(\lambda)$.

LEMMA 6.1.2. *Set*

$$\widetilde{m}^\sharp(\lambda):=\begin{cases}\widetilde{m}^\flat(\lambda)\prod\limits_{l=1}^{n-1}(d_{l_+}(\lambda))^{-\sigma_3}, & \lambda\in\mathbb{C}\setminus\left(\widehat{\Gamma}\cup\left(\bigcup\limits_{i=1}^{n-1}(\widetilde{K}_i^\pm\cup\widetilde{L}_i^\pm)\right)\right),\\ \widetilde{m}^\flat(\lambda)(\widetilde{J}_{\widetilde{K}_i^\pm}(\lambda))^{-1}\prod\limits_{l=1}^{n-1}(d_{l_-}(\lambda))^{-\sigma_3}, & \lambda\in\mathrm{int}\widetilde{K}_i^\pm,\quad 1\le i\le n-1,\\ \widetilde{m}^\flat(\lambda)(\widetilde{J}_{\widetilde{L}_i^\pm}(\lambda))^{-1}\prod\limits_{l=1}^{n-1}(d_{l_-}(\lambda))^{-\sigma_3}, & \lambda\in\mathrm{int}\widetilde{L}_i^\pm,\quad 1\le i\le n-1,\end{cases}$$

where

$$d_{l_+}(\lambda):=\tfrac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)},\quad \lambda\in\mathbb{C}\setminus\left(\bigcup_{i=1}^{n-1}(\widetilde{K}_i^\pm\cup\widetilde{L}_i^\pm)\right),\quad 1\le l\le n-1,$$

$$d_{l_-}(\lambda):=\begin{cases}\frac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda\pm\lambda_l)}, & \lambda\in\bigcup\limits_{i=1}^{n-1}\mathrm{int}\widetilde{K}_i^\pm,\quad 1\le l\le n-1,\\ \frac{(\lambda\pm\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)}, & \lambda\in\bigcup\limits_{i=1}^{n-1}\mathrm{int}\widetilde{L}_i^\pm,\quad 1\le l\le n-1,\end{cases}$$

and the $\mathrm{SL}(2,\mathbb{C})$*-valued, holomorphic in* $\mathrm{int}\widetilde{K}_i^\pm$ *and* $\mathrm{int}\widetilde{L}_i^\pm$*, resp., functions* $\widetilde{J}_{\widetilde{K}_i^\pm}(\lambda)$ *and* $\widetilde{J}_{\widetilde{L}_i^\pm}(\lambda)$, $1\le i\le n-1$, *are given by*

$$\widetilde{J}_{\widetilde{K}_i^\pm}(\lambda)=\begin{pmatrix}\dfrac{\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}^{-1}(\lambda)}-\frac{v_i(\widetilde{\delta}(\pm\lambda_i))^{-2}\widetilde{C}_i^\sharp}{(d_{i_-}(\lambda))^2}\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}(\lambda)}}{(\lambda\mp\lambda_i)} & \frac{\widetilde{C}_i^\sharp}{(d_{i_-}(\lambda))^2}\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}(\lambda)}\\ -v_i(\widetilde{\delta}(\pm\lambda_i))^{-2}\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)} & (\lambda\mp\lambda_i)\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)}\end{pmatrix},$$

$$\widetilde{J}_{\widetilde{L}_i^\pm}(\lambda)=\begin{pmatrix}(\lambda\mp\overline{\lambda_i})\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}^{-1}(\lambda)} & \overline{v_i}(\widetilde{\delta}(\pm\overline{\lambda_i}))^2\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}^{-1}(\lambda)}{d_{l_+}^{-1}(\lambda)}\\ -\frac{\overline{\widetilde{C}_i^\sharp}}{(d_{i_-}(\lambda))^{-2}}\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}(\lambda)}{d_{l_+}^{-1}(\lambda)} & \dfrac{\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}(\lambda)}{d_{l_+}(\lambda)}-\frac{\overline{v_i}(\widetilde{\delta}(\pm\overline{\lambda_i}))^2\overline{\widetilde{C}_i^\sharp}}{(d_{i_-}(\lambda))^{-2}}\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\frac{d_{l_-}(\lambda)}{d_{l_+}^{-1}(\lambda)}}{(\lambda\mp\overline{\lambda_i})}\end{pmatrix},$$

with

$$\widetilde{C}_i^\sharp=(v_i)^{-1}(\widetilde{\delta}(\pm\lambda_i))^2(d_{i_-}(\pm\lambda_i))^2\prod_{\substack{l=1\\ \neq i}}^{n-1}(d_{l_+}(\pm\lambda_i))^2,\quad 1\le i\le n-1.$$

Then $\widetilde{m}^\sharp(\lambda)\colon\mathbb{C}\setminus((\mathcal{Z}_d\setminus\cup_{i=1}^{n-1}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\widehat{\Gamma}\cup(\cup_{i=1}^{n-1}(\widetilde{K}_i^\pm\cup\widetilde{L}_i^\pm))))\to\mathrm{SL}(2,\mathbb{C})$ *solves the following, extended RH problem on* $(\sigma_{\mathcal{L}}\setminus\cup_{i=1}^{n-1}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\cup_{i=1}^{n-1}(\widetilde{K}_i^\pm\cup\widetilde{L}_i^\pm))$*:*

$$\widetilde{m}^\sharp_+(\lambda)=\widetilde{m}^\sharp_-(\lambda)e^{-i\theta(\lambda)\mathrm{ad}(\sigma_3)}\widetilde{v}^\sharp(\lambda),$$

where

$$\widetilde{v}^\sharp(\lambda)|_{\widehat{\Gamma}}=\begin{pmatrix}(1-r(\lambda)\overline{r(\overline{\lambda})})\frac{\widetilde{\delta}_-(\lambda)}{\widetilde{\delta}_+(\lambda)} & \frac{r(\lambda)}{(\widetilde{\delta}_-(\lambda)\widetilde{\delta}_+(\lambda))^{-1}}\prod\limits_{l=1}^{n-1}\left(\frac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)}\right)^2\\ -\frac{\overline{r(\overline{\lambda})}}{\widetilde{\delta}_-(\lambda)\widetilde{\delta}_+(\lambda)}\prod\limits_{l=1}^{n-1}\left(\frac{(\lambda-\overline{\lambda_l})(\lambda+\overline{\lambda_l})}{(\lambda-\lambda_l)(\lambda+\lambda_l)}\right)^{-2} & \frac{\widetilde{\delta}_+(\lambda)}{\widetilde{\delta}_-(\lambda)}\end{pmatrix},$$

$$\widetilde{v}^{\sharp}(\lambda)=\begin{cases}\mathrm{I}+\frac{(v_i)^{-1}(\widetilde{\delta}(\pm\lambda_i))^2}{(\lambda\mp\lambda_i)}\left(\frac{\lambda_i^2-\overline{\lambda_i}^2}{2\lambda_i}\right)^2\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\left(\frac{\overline{\lambda_l}^2-\lambda_i^2}{\lambda_l^2-\lambda_i^2}\right)^2\sigma_+, & \lambda\in\bigcup\limits_{i=1}^{n-1}\widetilde{K}_i^{\pm},\\ \mathrm{I}+\frac{(\overline{v_i})^{-1}(\widetilde{\delta}(\pm\overline{\lambda_i}))^{-2}}{(\lambda\mp\overline{\lambda_i})}\left(\frac{\lambda_i^2-\overline{\lambda_i}^2}{2\overline{\lambda_i}}\right)^2\prod\limits_{\substack{l=1\\ \neq i}}^{n-1}\left(\frac{\lambda_l^2-\overline{\lambda_i}^2}{\overline{\lambda_l}^2-\overline{\lambda_i}^2}\right)^2\sigma_-, & \lambda\in\bigcup\limits_{i=1}^{n-1}\widetilde{L}_i^{\pm},\end{cases}$$

with polar (residue) conditions

$$\begin{aligned}\mathrm{res}(\widetilde{m}^{\sharp}(\lambda);\lambda_i)&=\lim_{\lambda\to\lambda_i}\widetilde{m}^{\sharp}(\lambda)v_i(\widetilde{\delta}(\lambda_i))^{-2}\prod_{l=1}^{n-1}\left(\frac{(\lambda_i-\lambda_l)(\lambda_i+\lambda_l)}{(\lambda_i-\overline{\lambda_l})(\lambda_i+\overline{\lambda_l})}\right)^2\sigma_-,\quad n\le i\le N,\\ \mathrm{res}(\widetilde{m}^{\sharp}(\lambda);-\lambda_i)&=-\sigma_3\mathrm{res}(\widetilde{m}^{\sharp}(\lambda);\lambda_i)\sigma_3,\quad n\le i\le N,\\ \mathrm{res}(\widetilde{m}^{\sharp}(\lambda);\overline{\lambda_i})&=\lim_{\lambda\to\overline{\lambda_i}}\widetilde{m}^{\sharp}(\lambda)\overline{v_i}(\widetilde{\delta}(\overline{\lambda_i}))^2\prod_{l=1}^{n-1}\left(\frac{(\overline{\lambda_i}-\overline{\lambda_l})(\overline{\lambda_i}+\overline{\lambda_l})}{(\overline{\lambda_i}-\lambda_l)(\overline{\lambda_i}+\lambda_l)}\right)^2\sigma_+,\quad n\le i\le N,\\ \mathrm{res}(\widetilde{m}^{\sharp}(\lambda);-\overline{\lambda_i})&=-\sigma_3\mathrm{res}(\widetilde{m}^{\sharp}(\lambda);\overline{\lambda_i})\sigma_3,\quad n\le i\le N,\end{aligned}$$

and, as $\lambda\to\infty$, $\lambda\in\mathbb{C}\backslash((\mathcal{Z}_d\backslash\cup_{i=1}^{n-1}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\}))\cup(\widehat{\Gamma}\cup(\cup_{i=1}^{n-1}(\widetilde{K}_i^{\pm}\cup\widetilde{L}_i^{\pm}))))$,

$$\widetilde{m}^{\sharp}(\lambda)=\mathrm{I}+\mathcal{O}(\lambda^{-1});$$

moreover, $Q(x,t)=2i\lim_{\lambda\to\infty}(\lambda\widetilde{m}^{\sharp}(x,t;\lambda))_{12}$ *is equal to* $Q(x,t)$ *in Lemma* 2.2, (11).

LEMMA 6.1.3. *Let* $\widetilde{\chi}(\lambda)$ *solve the following RH problem on* $\sigma_{\mathcal{L}}\backslash\cup_{i=1}^{n-1}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})$:

$$\widetilde{\chi}_+(\lambda)=\widetilde{\chi}_-(\lambda)e^{-i\theta(\lambda)\mathrm{ad}(\sigma_3)}\widetilde{v}^{\sharp}(\lambda)|_{\widehat{\Gamma}},\quad \lambda\in\widehat{\Gamma},$$

with polar (residue) conditions

$$\begin{aligned}\mathrm{res}(\widetilde{\chi}(\lambda);\lambda_i)&=\lim_{\lambda\to\lambda_i}\widetilde{\chi}(\lambda)v_i(\widetilde{\delta}(\lambda_i))^{-2}\prod_{l=1}^{n-1}\left(\frac{(\lambda_i-\lambda_l)(\lambda_i+\lambda_l)}{(\lambda_i-\overline{\lambda_l})(\lambda_i+\overline{\lambda_l})}\right)^2\sigma_-,\quad n\le i\le N,\\ \mathrm{res}(\widetilde{\chi}(\lambda);-\lambda_i)&=-\sigma_3\mathrm{res}(\widetilde{\chi}(\lambda);\lambda_i)\sigma_3,\quad n\le i\le N,\\ \mathrm{res}(\widetilde{\chi}(\lambda);\overline{\lambda_i})&=\lim_{\lambda\to\overline{\lambda_i}}\widetilde{\chi}(\lambda)\overline{v_i}(\widetilde{\delta}(\overline{\lambda_i}))^2\prod_{l=1}^{n-1}\left(\frac{(\overline{\lambda_i}-\overline{\lambda_l})(\overline{\lambda_i}+\overline{\lambda_l})}{(\overline{\lambda_i}-\lambda_l)(\overline{\lambda_i}+\lambda_l)}\right)^2\sigma_+,\quad n\le i\le N,\\ \mathrm{res}(\widetilde{\chi}(\lambda);-\overline{\lambda_i})&=-\sigma_3\mathrm{res}(\widetilde{\chi}(\lambda);\overline{\lambda_i})\sigma_3,\quad n\le i\le N,\end{aligned}$$

and, as $\lambda\to\infty$, $\lambda\in\mathbb{C}\backslash(\cup_{i=n}^{N}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})\cup\widehat{\Gamma})$,

$$\widetilde{\chi}(\lambda)=\mathrm{I}+\mathcal{O}(\lambda^{-1}).$$

Then as $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$ *and* $(x,t)\in\Omega_n$, *the function* $\widetilde{E}(\lambda):=\widetilde{m}^{\sharp}(\lambda)(\widetilde{\chi}(\lambda))^{-1}$ *has the following asymptotics:*

$$\widetilde{E}(\lambda)=\mathrm{I}+\mathcal{O}(\widetilde{F}(\lambda;\lambda_0)\exp\{\widetilde{a}\widetilde{b}t\}),$$

where $||\widetilde{F}(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C};M_2(\mathbb{C}))}<\infty$, $||\widetilde{F}(\lambda;\cdot)||_{\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{C}))}<\infty$, $\widetilde{F}(\lambda;\lambda_0)\sim\mathcal{O}\left(\frac{C(\lambda_0)}{\lambda}\right)$ *as* $\lambda\to\infty$ *with* $C(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};M_2(\mathbb{C}))$, $\widetilde{a}:=8\min\{\eta_i\}_{i=1}^{n-1}\,(>0)$, *and* $\widetilde{b}:=\min\{|\xi_n-\xi_i|\}_{i=1}^{n-1}$.

6.2. Asymptotic solution for $\widetilde{\chi}(\lambda)$.

PROPOSITION 6.2.1. *The solution of the model RH problem formulated in Lemma* 6.1.3, $\widetilde{\chi}(\lambda)\colon \mathbb{C}\setminus(\widehat{\Gamma}\cup(\cup_{i=n}^{N}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})))\to \mathrm{SL}(2,\mathbb{C})$, *has the following representation:*

$$\widetilde{\chi}(\lambda)=\widetilde{\chi}_d(\lambda)+\int_{\widehat{\Gamma}}\frac{\widetilde{\chi}_-(\varrho)(\widetilde{v}^{\sharp}(\varrho)|_{\widehat{\Gamma}}-\mathrm{I})}{(\varrho-\lambda)}\frac{d\varrho}{2\pi i}, \tag{107}$$

where

$$\begin{aligned}\widetilde{\chi}_d(\lambda)&=\mathrm{I}\\ &+\sum_{i=n}^{N}\left(\frac{\mathrm{res}(\widetilde{\chi}(\lambda);\lambda_i)}{(\lambda-\lambda_i)}-\frac{\sigma_3\mathrm{res}(\widetilde{\chi}(\lambda);\lambda_i)\sigma_3}{(\lambda+\lambda_i)}+\frac{\mathrm{res}(\widetilde{\chi}(\lambda);\overline{\lambda_i})}{(\lambda-\overline{\lambda_i})}-\frac{\sigma_3\mathrm{res}(\widetilde{\chi}(\lambda);\overline{\lambda_i})\sigma_3}{(\lambda+\overline{\lambda_i})}\right).\end{aligned} \tag{108}$$

The solution of (107) *can be written as the following ordered product:*

$$\widetilde{\chi}(\lambda)=\widetilde{\chi}_d(\lambda)\widetilde{\chi}^c(\lambda),$$

where $\widetilde{\chi}_d(\lambda)$ *is given by* (108), *and* $\widetilde{\chi}^c(\lambda)$ *solves the following RH problem:* (1) $\widetilde{\chi}^c(\lambda)$ *is piecewise holomorphic for all* $\lambda\in\mathbb{C}\setminus\widehat{\Gamma}$; (2) $\widetilde{\chi}^c_+(\lambda)=\widetilde{\chi}^c_-(\lambda)\exp\{-i\theta(\lambda)\mathrm{ad}(\sigma_3)\}(\widetilde{v}^{\sharp}(\lambda)|_{\widehat{\Gamma}})$, $\lambda\in\widehat{\Gamma}$; *and* (3) *as* $\lambda\to\infty$, $\lambda\in\mathbb{C}\setminus\widehat{\Gamma}$, $\widetilde{\chi}^c(\lambda)=\mathrm{I}+\mathcal{O}(\lambda^{-1})$.

LEMMA 6.2.1. *Let* $\widetilde{\epsilon}_0$ *denote an arbitrarily fixed, sufficiently small positive real number. For* $\widetilde{\aleph}\in\{0,\pm\lambda_0\}$, *set* $\widetilde{\mathcal{N}}(\widetilde{\aleph};\widetilde{\epsilon}_0):=\{\lambda;\ |\lambda-\widetilde{\aleph}|\le\widetilde{\epsilon}_0\}$. *Then as* $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$ *and* $\lambda\in\mathbb{C}\setminus\cup_{\widetilde{\aleph}\in\{0,\pm\lambda_0\}}\widetilde{\mathcal{N}}(\widetilde{\aleph};\widetilde{\epsilon}_0)$, $\widetilde{\chi}^c(\lambda)$ *has the following asymptotic expansion:*

$$\begin{aligned}\widetilde{\chi}^c(\lambda)=\mathrm{I}+\frac{1}{4}\sqrt{-\frac{\nu(\lambda_0)}{2\lambda_0^2t}}\left(\frac{1}{\lambda-\lambda_0}+\frac{1}{\lambda+\lambda_0}\right)\Big(&\exp\{-i(\phi^-(\lambda_0)+\widehat{\Phi}^-(\lambda_0;t))\}\sigma_-\\ &+\ \exp\{i(\phi^-(\lambda_0)+\widehat{\Phi}^-(\lambda_0;t))\}\sigma_+\Big)+\mathcal{O}\left(\frac{\widetilde{G}(\lambda;\lambda_0)\ln|t|}{t}\right),\end{aligned}$$

where $\nu(\lambda_0)$, $\phi^-(\lambda_0)$, *and* $\widehat{\Phi}^-(\lambda_0;t)$ *are given in Theorem* 2.1, *equations* (21), (23), *and* (24), $||\widetilde{G}(\cdot;\lambda_0)||_{\mathcal{L}^\infty(\mathbb{C}\setminus\cup_{\widetilde{\aleph}\in\{0,\pm\lambda_0\}}\widetilde{\mathcal{N}}(\widetilde{\aleph};\widetilde{\epsilon}_0);M_2(\mathbb{C}))}<\infty$, $\widetilde{G}(\lambda;\cdot)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{C}))$, $\widetilde{G}(\lambda;\lambda_0)\sim\mathcal{O}\left(\frac{C(\lambda_0)}{\lambda}\right)$ *as* $\lambda\to\infty$ *with* $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};M_2(\mathbb{C}))$, *and satisfies the following involutions,* $\widetilde{\chi}^c(-\lambda)=\sigma_3\widetilde{\chi}^c(\lambda)\sigma_3$ *and* $\widetilde{\chi}^c(\lambda)=\sigma_1\overline{\widetilde{\chi}^c(\overline{\lambda})}\sigma_1$.

PROPOSITION 6.2.2. *For* $n\le i\le N$, *set*

$$\mathrm{res}(\widetilde{\chi}(\lambda);\lambda_i)=\begin{pmatrix}\alpha_i^- & a_i^-\\ \beta_i^- & b_i^-\end{pmatrix},\qquad \mathrm{res}(\widetilde{\chi}(\lambda);\overline{\lambda_i})=\begin{pmatrix}c_i^- & \omega_i^-\\ d_i^- & \widehat{\delta}_i^-\end{pmatrix}.$$

Then as $\lambda\to\infty$, $\lambda\in\mathbb{C}\setminus(\widehat{\Gamma}\cup(\cup_{i=n}^{N}(\{\pm\lambda_i\}\cup\{\pm\overline{\lambda_i}\})))$, $\widetilde{\chi}(\lambda)$ *has the following asymptotic expansion,*

$$\begin{aligned}\widetilde{\chi}(\lambda)=\mathrm{I}+\frac{1}{2\lambda}\Bigg(&\left\{\overline{Q^{\widetilde{\chi}}(x,t)}+4\sum_{i=n}^{N}\left(\beta_i^--\frac{\widetilde{\chi}^c_{21}(\overline{\lambda_i})}{\widetilde{\chi}^c_{11}(\overline{\lambda_i})}\widehat{\delta}_i^-\right)\right\}\sigma_-\\ &+\ \left\{Q^{\widetilde{\chi}}(x,t)+4\sum_{i=n}^{N}\left(\omega_i^--\frac{\widetilde{\chi}^c_{12}(\lambda_i)}{\widetilde{\chi}^c_{22}(\lambda_i)}\alpha_i^-\right)\right\}\sigma_+\Bigg)+\mathcal{O}(\lambda^{-2}),\end{aligned}$$

where $\lim_{\lambda\to\infty}(\widetilde{\chi}^c(x,t;\lambda))_{12} := \widetilde{Q^{\chi}}(x,t)/2\lambda$, $\{\alpha_i^-, \omega_i^-\}_{i=n}^N$ *satisfy the following nondegenerate system of* $2(N-n+1)$ *linear inhomogeneous algebraic equations,*

$$\begin{bmatrix} \boxed{\widehat{\mathcal{A}}^-} & \boxed{\widehat{\mathcal{B}}^-} \\ \boxed{\widehat{\mathcal{C}}^-} & \boxed{\widehat{\mathcal{D}}^-} \end{bmatrix} \begin{bmatrix} \alpha_n^- \\ \alpha_{n+1}^- \\ \vdots \\ \alpha_N^- \\ \omega_n^- \\ \omega_{n+1}^- \\ \vdots \\ \omega_N^- \end{bmatrix} = \begin{bmatrix} g_n^- \widetilde{\chi}_{12}^c(\lambda_n) \\ g_{n+1}^- \widetilde{\chi}_{12}^c(\lambda_{n+1}) \\ \vdots \\ g_N^- \widetilde{\chi}_{12}^c(\lambda_N) \\ \overline{g_n^-}\, \widetilde{\chi}_{11}^c(\overline{\lambda_n}) \\ \overline{g_{n+1}^-}\, \widetilde{\chi}_{11}^c(\overline{\lambda_{n+1}}) \\ \vdots \\ \overline{g_N^-}\, \widetilde{\chi}_{11}^c(\overline{\lambda_N}) \end{bmatrix},$$

where, for $i,j\in\{n,n+1,\dots,N\}$, *the* $(N-n+1)\times(N-n+1)$ *matrix blocks* $\widehat{\mathcal{A}}^-$, $\widehat{\mathcal{B}}^-$, $\widehat{\mathcal{C}}^-$, *and* $\widehat{\mathcal{D}}^-$ *are defined as follows,*

$$\widehat{\mathcal{A}}_{ij}^- := \begin{cases} \frac{\lambda_i + g_i^- \widetilde{\chi}_{12}^c(\lambda_i)\widetilde{\chi}_{22}^c(\lambda_i) + \lambda_i g_i^- \mathrm{W}(\widetilde{\chi}_{12}^c(\lambda_i), \widetilde{\chi}_{22}^c(\lambda_i))}{\lambda_i \widetilde{\chi}_{22}^c(\lambda_i)}, & i=j, \\ -\frac{2g_i^-(-\lambda_i \widetilde{\chi}_{22}^c(\lambda_i)\widetilde{\chi}_{12}^c(\lambda_j) + \lambda_j \widetilde{\chi}_{22}^c(\lambda_j)\widetilde{\chi}_{12}^c(\lambda_i))}{\widetilde{\chi}_{22}^c(\lambda_j)(\lambda_i^2-\lambda_j^2)}, & i\neq j, \end{cases}$$

$$\widehat{\mathcal{B}}_{ij}^- := \begin{cases} -\frac{2g_i^-(\lambda_i \widetilde{\chi}_{22}^c(\lambda_i)\widetilde{\chi}_{11}^c(\overline{\lambda_i}) - \overline{\lambda_i}\widetilde{\chi}_{21}^c(\overline{\lambda_i})\widetilde{\chi}_{12}^c(\lambda_i))}{\widetilde{\chi}_{11}^c(\overline{\lambda_i})(\lambda_i^2-\overline{\lambda_i}^2)}, & i=j, \\ -\frac{2g_i^-(\lambda_i \widetilde{\chi}_{22}^c(\lambda_i)\widetilde{\chi}_{11}^c(\overline{\lambda_j}) - \overline{\lambda_j}\widetilde{\chi}_{21}^c(\overline{\lambda_j})\widetilde{\chi}_{12}^c(\lambda_i))}{\widetilde{\chi}_{11}^c(\overline{\lambda_j})(\lambda_i^2-\overline{\lambda_j}^2)}, & i\neq j, \end{cases}$$

$$\widehat{\mathcal{C}}_{ij}^- := \begin{cases} -\frac{2\overline{g_i^-}(\lambda_i \widetilde{\chi}_{22}^c(\lambda_i)\widetilde{\chi}_{11}^c(\overline{\lambda_i}) - \overline{\lambda_i}\widetilde{\chi}_{21}^c(\overline{\lambda_i})\widetilde{\chi}_{12}^c(\lambda_i))}{\widetilde{\chi}_{22}^c(\lambda_i)(\overline{\lambda_i}^2-\lambda_i^2)}, & i=j, \\ -\frac{2\overline{g_i^-}(-\overline{\lambda_i}\widetilde{\chi}_{21}^c(\overline{\lambda_i})\widetilde{\chi}_{12}^c(\lambda_j) + \lambda_j \widetilde{\chi}_{22}^c(\lambda_j)\widetilde{\chi}_{11}^c(\overline{\lambda_i}))}{\widetilde{\chi}_{22}^c(\lambda_j)(\overline{\lambda_i}^2-\lambda_j^2)}, & i\neq j, \end{cases}$$

$$\widehat{\mathcal{D}}_{ij}^- := \begin{cases} \frac{\overline{\lambda_i} - \overline{g_i^-}\widetilde{\chi}_{21}^c(\overline{\lambda_i})\widetilde{\chi}_{11}^c(\overline{\lambda_i}) + \overline{\lambda_i}\,\overline{g_i^-}\mathrm{W}(\widetilde{\chi}_{21}^c(\overline{\lambda_i}), \widetilde{\chi}_{11}^c(\overline{\lambda_i}))}{\overline{\lambda_i}\widetilde{\chi}_{11}^c(\overline{\lambda_i})}, & i=j, \\ -\frac{2\overline{g_i^-}(\overline{\lambda_i}\widetilde{\chi}_{21}^c(\overline{\lambda_i})\widetilde{\chi}_{11}^c(\overline{\lambda_j}) - \overline{\lambda_j}\widetilde{\chi}_{21}^c(\overline{\lambda_j})\widetilde{\chi}_{11}^c(\overline{\lambda_i}))}{\widetilde{\chi}_{11}^c(\overline{\lambda_j})(\overline{\lambda_i}^2-\overline{\lambda_j}^2)}, & i\neq j, \end{cases}$$

$\{\beta_i^-, \widehat{\delta}_i^-\}_{i=n}^N$ *satisfy the following nondegenerate system of* $2(N-n+1)$ *linear inhomogeneous algebraic equations,*

$$\begin{bmatrix} \boxed{\widehat{\mathcal{E}}^-} & \boxed{\widehat{\mathcal{F}}^-} \\ \boxed{\widehat{\mathcal{G}}^-} & \boxed{\widehat{\mathcal{H}}^-} \end{bmatrix} \begin{bmatrix} \beta_n^- \\ \beta_{n+1}^- \\ \vdots \\ \beta_N^- \\ \widehat{\delta}_n^- \\ \widehat{\delta}_{n+1}^- \\ \vdots \\ \widehat{\delta}_N^- \end{bmatrix} = \begin{bmatrix} g_n^- \widetilde{\chi}_{22}^c(\lambda_n) \\ g_{n+1}^- \widetilde{\chi}_{22}^c(\lambda_{n+1}) \\ \vdots \\ g_N^- \widetilde{\chi}_{22}^c(\lambda_N) \\ \overline{g_n^-}\, \widetilde{\chi}_{21}^c(\overline{\lambda_n}) \\ \overline{g_{n+1}^-}\, \widetilde{\chi}_{21}^c(\overline{\lambda_{n+1}}) \\ \vdots \\ \overline{g_N^-}\, \widetilde{\chi}_{21}^c(\overline{\lambda_N}) \end{bmatrix},$$

where, for $i,j\in\{n,n+1,\dots,N\}$, *the* $(N-n+1)\times(N-n+1)$ *matrix blocks* $\widehat{\mathcal{E}}^-$, $\widehat{\mathcal{F}}^-$, $\widehat{\mathcal{G}}^-$, *and* $\widehat{\mathcal{H}}^-$ *are defined as follows,*

$$\widehat{\mathcal{E}}^-_{ij}:=\begin{cases}\frac{\lambda_i-g_i^-\widetilde{\chi}^c_{12}(\lambda_i)\widetilde{\chi}^c_{22}(\lambda_i)+\lambda_i g_i^-\mathrm{W}(\widetilde{\chi}^c_{12}(\lambda_i),\widetilde{\chi}^c_{22}(\lambda_i))}{\lambda_i\widetilde{\chi}^c_{22}(\lambda_i)}, & i=j,\\ \frac{2g_i^-(\lambda_j\widetilde{\chi}^c_{12}(\lambda_j)\widetilde{\chi}^c_{22}(\lambda_i)-\lambda_i\widetilde{\chi}^c_{12}(\lambda_i)\widetilde{\chi}^c_{22}(\lambda_j))}{\widetilde{\chi}^c_{22}(\lambda_j)(\lambda_i^2-\lambda_j^2)}, & i\neq j,\end{cases}$$

$$\widehat{\mathcal{F}}^-_{ij}:=\begin{cases}\frac{2g_i^-(\lambda_i\widetilde{\chi}^c_{12}(\lambda_i)\widetilde{\chi}^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\widetilde{\chi}^c_{11}(\overline{\lambda_i})\widetilde{\chi}^c_{22}(\lambda_i))}{\widetilde{\chi}^c_{11}(\overline{\lambda_i})(\lambda_i^2-\overline{\lambda_i}^2)}, & i=j,\\ \frac{2g_i^-(\lambda_i\widetilde{\chi}^c_{12}(\lambda_i)\widetilde{\chi}^c_{21}(\overline{\lambda_j})-\overline{\lambda_j}\widetilde{\chi}^c_{11}(\overline{\lambda_j})\widetilde{\chi}^c_{22}(\lambda_i))}{\widetilde{\chi}^c_{11}(\overline{\lambda_j})(\lambda_i^2-\overline{\lambda_j}^2)}, & i\neq j,\end{cases}$$

$$\widehat{\mathcal{G}}^-_{ij}:=\begin{cases}\frac{2\overline{g_i^-}(\lambda_i\widetilde{\chi}^c_{12}(\lambda_i)\widetilde{\chi}^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\widetilde{\chi}^c_{11}(\overline{\lambda_i})\widetilde{\chi}^c_{22}(\lambda_i))}{\widetilde{\chi}^c_{22}(\lambda_i)(\overline{\lambda_i}^2-\lambda_i^2)}, & i=j,\\ \frac{2\overline{g_i^-}(\lambda_j\widetilde{\chi}^c_{12}(\lambda_j)\widetilde{\chi}^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\widetilde{\chi}^c_{11}(\overline{\lambda_i})\widetilde{\chi}^c_{22}(\lambda_j))}{\widetilde{\chi}^c_{22}(\lambda_j)(\overline{\lambda_i}^2-\lambda_j^2)}, & i\neq j,\end{cases}$$

$$\widehat{\mathcal{H}}^-_{ij}:=\begin{cases}\frac{\overline{\lambda_i}+\overline{g_i^-}\widetilde{\chi}^c_{11}(\overline{\lambda_i})\widetilde{\chi}^c_{21}(\overline{\lambda_i})-\overline{\lambda_i}\,\overline{g_i^-}\mathrm{W}(\widetilde{\chi}^c_{11}(\overline{\lambda_i}),\widetilde{\chi}^c_{21}(\overline{\lambda_i}))}{\overline{\lambda_i}\,\widetilde{\chi}^c_{11}(\overline{\lambda_i})}, & i=j,\\ \frac{2\overline{g_i^-}(\overline{\lambda_i}\widetilde{\chi}^c_{11}(\overline{\lambda_i})\widetilde{\chi}^c_{21}(\overline{\lambda_j})-\overline{\lambda_j}\widetilde{\chi}^c_{11}(\overline{\lambda_j})\widetilde{\chi}^c_{21}(\overline{\lambda_i}))}{\widetilde{\chi}^c_{11}(\overline{\lambda_j})(\overline{\lambda_i}^2-\overline{\lambda_j}^2)}, & i\neq j,\end{cases}$$

with $a_i^-=-\frac{\widetilde{\chi}^c_{12}(\lambda_i)}{\widetilde{\chi}^c_{22}(\lambda_i)}\alpha_i^-$, $b_i^-=-\frac{\widetilde{\chi}^c_{12}(\lambda_i)}{\widetilde{\chi}^c_{22}(\lambda_i)}\beta_i^-$, $c_i^-=-\frac{\widetilde{\chi}^c_{21}(\overline{\lambda_i})}{\widetilde{\chi}^c_{11}(\overline{\lambda_i})}\omega_i^-$, $d_i^-=-\frac{\widetilde{\chi}^c_{21}(\overline{\lambda_i})}{\widetilde{\chi}^c_{11}(\overline{\lambda_i})}\widehat{\delta}_i^-$, $n\le i\le N$,

$$g_j^-:=C_je^{2i\lambda_j^2x+4i\lambda_j^4t}(\delta^-(\lambda_j;\lambda_0))^{-2}\prod_{l=1}^{n-1}\left(\frac{(\lambda_j-\lambda_l)(\lambda_j+\lambda_l)}{(\lambda_j-\overline{\lambda_l})(\lambda_j+\overline{\lambda_l})}\right)^2,\qquad n\le j\le N,$$

$\delta^-(\lambda_k;\lambda_0)$, $k\in\{n,n+1,\dots,N\}$, *given in Theorem* 2.1, (19), *and* $\mathrm{W}(\widetilde{\chi}^c_{ij}(z),\widetilde{\chi}^c_{i'j'}(z))$ *is the Wronskian of* $\widetilde{\chi}^c_{ij}(\lambda)$ *and* $\widetilde{\chi}^c_{i'j'}(\lambda)$ $(i,j,i',j'\in\{1,2\})$ *evaluated at* z : $\mathrm{W}(\widetilde{\chi}^c_{ij}(z),\widetilde{\chi}^c_{i'j'}(z)):=(\widetilde{\chi}^c_{ij}(\lambda)\partial_\lambda\widetilde{\chi}^c_{i'j'}(\lambda)-\widetilde{\chi}^c_{i'j'}(\lambda)\partial_\lambda\chi^c_{ij}(\lambda))|_{\lambda=z}$.

Corollary 6.2.1. *As* $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$ *and* $(x,t)\in\Omega_n$,

$$Q(x,t)=\underset{\sim}{Q}{}^{\mathcal{C}}_-(x,t)+4i\sum_{j=n}^N\left(\omega_j^--\frac{\widetilde{\chi}^c_{12}(\lambda_j)}{\widetilde{\chi}^c_{22}(\lambda_j)}\alpha_j^-\right)+\mathcal{O}(C(\lambda_0)\exp\{\widetilde{a}\widetilde{b}t\}),$$

where $\underset{\sim}{Q}{}^{\mathcal{C}}_-(x,t):=iQ^{\widetilde{\chi}}(x,t)$, $\widetilde{a}$ *and* $\widetilde{b}$ *are given in Lemma* 6.1.3, *and* $C(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{C})$.

Proposition 6.2.3. *As* $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$ *and* $(x,t)\in\Omega_n$,

$$Q(x,t)=Q^{\mathcal{C}}_-(x,t)+4i\left(\omega_n^--\frac{\widetilde{\chi}^c_{12}(\lambda_n)}{\widetilde{\chi}^c_{22}(\lambda_n)}\alpha_n^-\right)+\mathcal{O}\left(\frac{C_1(\lambda_0)\ln|t|}{t}\right)+\mathcal{O}(C_2(\lambda_0)e^{\widetilde{a}_0\widetilde{b}_0t}),$$

where $\alpha_i^-,\omega_i^-\sim\mathcal{O}(\exp\{\widetilde{a}^\flat\min_{n+1\le l\le N}|\xi_n-\xi_l|t\})$, $n+1\le i\le N$, $\widetilde{a}^\flat:=8\min\{\eta_l\}_{l=n+1}^N$ (>0),

$$\alpha_n^-=\frac{\widehat{a}^-_{12}\overline{g_n^-}\widetilde{\chi}^c_{11}(\overline{\lambda_n})+\widehat{a}^-_{22}g_n^-\widetilde{\chi}^c_{12}(\lambda_n)}{(\widehat{a}^-_{11}\widehat{a}^-_{22}-\widehat{a}^-_{12}\widehat{a}^-_{21})},$$

$$\omega_n^-=\frac{\widehat{a}^-_{11}\overline{g_n^-}\widetilde{\chi}^c_{11}(\overline{\lambda_n})+\widehat{a}^-_{21}g_n^-\widetilde{\chi}^c_{12}(\lambda_n)}{(\widehat{a}^-_{11}\widehat{a}^-_{22}-\widehat{a}^-_{12}\widehat{a}^-_{21})},$$

$$\widehat{a}_{11}^{-} := \frac{\lambda_n + g_n^- \widetilde{\chi}_{12}^c(\lambda_n)\widetilde{\chi}_{22}^c(\lambda_n) + \lambda_n g_n^- \mathrm{W}(\widetilde{\chi}_{12}^c(\lambda_n), \widetilde{\chi}_{22}^c(\lambda_n))}{\lambda_n \widetilde{\chi}_{22}^c(\lambda_n)},$$

$$\widehat{a}_{12}^{-} := \frac{2g_n^-(\lambda_n \widetilde{\chi}_{22}^c(\lambda_n)\widetilde{\chi}_{11}^c(\overline{\lambda_n}) - \overline{\lambda_n}\widetilde{\chi}_{21}^c(\overline{\lambda_n})\widetilde{\chi}_{12}^c(\lambda_n))}{\widetilde{\chi}_{11}^c(\overline{\lambda_n})(\lambda_n^2 - \overline{\lambda_n}^2)},$$

$$\widehat{a}_{21}^{-} := \frac{2\overline{g_n^-}(\lambda_n \widetilde{\chi}_{22}^c(\lambda_n)\widetilde{\chi}_{11}^c(\overline{\lambda_n}) - \overline{\lambda_n}\widetilde{\chi}_{21}^c(\overline{\lambda_n})\widetilde{\chi}_{12}^c(\lambda_n))}{\widetilde{\chi}_{22}^c(\lambda_n)(\overline{\lambda_n}^2 - \lambda_n^2)},$$

$$\widehat{a}_{22}^{-} := \frac{\overline{\lambda_n} - \overline{g_n^-}\widetilde{\chi}_{21}^c(\overline{\lambda_n})\widetilde{\chi}_{11}^c(\overline{\lambda_n}) + \overline{\lambda_n}\,\overline{g_n^-}\mathrm{W}(\widetilde{\chi}_{21}^c(\overline{\lambda_n}), \widetilde{\chi}_{11}^c(\overline{\lambda_n}))}{\overline{\lambda_n}\widetilde{\chi}_{11}^c(\overline{\lambda_n})},$$

$Q_-^{\mathcal{C}}(x,t)$ is given in Theorem 2.1, *equations* (20), (21), (23), *and* (24), $\widetilde{a}_0 := \min(\widetilde{a}, \widetilde{a}^\flat)$ (> 0), $\widetilde{b}_0 := \min(\widetilde{b}, \min\{|\xi_n - \xi_l|\}_{l=n+1}^N)$, $C_1(\lambda_0) \in \mathcal{S}(\mathbb{R}_{>M}; \mathbb{C})$, *and* $C_2(\lambda_0) \in \mathcal{L}^\infty(\mathbb{R}_{>M}; \mathbb{C})$.

PROPOSITION 6.2.4. *As $t \to -\infty$ and $x \to +\infty$ such that $\lambda_0 > M$ and $(x,t) \in \Omega_n$,*

$$Q(x,t) = Q_{\mathrm{as}}^-(x,t) + \mathcal{O}\left(\frac{C(\lambda_0)\ln|t|}{t}\right),$$

where $Q_{\mathrm{as}}^-(x,t)$ is given in Theorem 2.1, (14)–(29), *and $C(\lambda_0) \in \mathcal{S}(\mathbb{R}_{>M}; \mathbb{C})$.*

6.3. Asymptotics of $((\boldsymbol{\Psi}^{-1}(\boldsymbol{x},\boldsymbol{t};\boldsymbol{0}))_{11})^2$ as $t \to -\infty$.

PROPOSITION 6.3.1. *Define $Q^\natural(x,t) := 2i \lim_{\lambda\to\infty}(\lambda \widetilde{\chi}^c(x,t;\lambda))_{12}$. Then*

$$(\widetilde{\chi}_{22}^c(0))^2 = (\delta^-(0;\lambda_0))^2 \exp\left\{ i \int_{+\infty}^x |Q^\natural(\varrho,t)|^2 d\varrho \right\}.$$

LEMMA 6.3.1 (see [35]). *As $t \to -\infty$ and $x \to +\infty$ such that $\lambda_0 > M$,*

$$Q^\natural(x,t) = \frac{u_{1,1,0}^-(\lambda_0) e^{i(4\lambda_0^4 t + \nu(\lambda_0)\ln|t|)}}{\sqrt{-t}} + \frac{u_{-1,2,0}^-(\lambda_0)}{(-t)} + \mathcal{O}\left(\frac{C(\lambda_0)(\ln|t|)^2}{(-t)^{3/2}}\right),$$

where

$$u_{1,1,0}^-(\lambda_0) = \sqrt{\frac{\nu(\lambda_0)}{2\lambda_0^2}} \exp\{i\theta^-(\lambda_0)\},$$

$$\theta^-(\lambda_0) = \phi^-(\lambda_0) + \frac{3\pi}{4} - \arg\Gamma(i\nu(\lambda_0)) + \arg r(\lambda_0) + 3\nu(\lambda_0)\ln 2 + 2\sum_{l=1}^{n-1} \arg\left(\frac{(\lambda_0 - \overline{\lambda_l})(\lambda_0 + \overline{\lambda_l})}{(\lambda_0 - \lambda_l)(\lambda_0 + \lambda_l)}\right),$$

$$u_{-1,2,0}^-(\lambda_0) = \frac{i}{8\pi\lambda_0^2}\left(\left.\frac{d(r(\varrho)|_{\varrho\in\mathbb{R}})}{d\varrho}\right|_{\varrho=0} - \left.\frac{d(r(\varrho)|_{\varrho\in i\mathbb{R}})}{d\varrho}\right|_{\varrho=0} \right) \times \exp\left\{ i\left(4\sum_{l=1}^{n-1}\gamma_l + 2\vartheta^-(\lambda_0) \right) \right\},$$

$$\vartheta^-(\lambda_0) = -\int_{\lambda_0}^\infty \frac{\ln(1 - |r(\varrho)|^2)}{\varrho} \frac{d\varrho}{\pi},$$

$\phi^-(\cdot)$ is given in Theorem 2.1, (23), *and* $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

PROPOSITION 6.3.2. *As* $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$,

$$(\widetilde{\chi}^c_{22}(0))^2=\exp\left\{i\left(\sqrt{\frac{2}{-t}}\int_{\lambda_0}^{\infty}\frac{\sqrt{\nu(\mu)}}{\mu^2}\big(R_i^-(0)\cos(\kappa^-(\mu;t))-R_r^-(0)\sin(\kappa^-(\mu;t))\big)\frac{d\mu}{\pi}\right)\right\}$$
$$+\mathcal{O}\left(\frac{C(\lambda_0)(\ln|t|)^2}{\lambda_0^2 t}\right),$$

where

$$R_i^-(0)\,=\,\Im\{R^-(0)\},\qquad R_r^-(0)=\Re\{R^-(0)\},$$
$$R^-(0):=\left(\frac{d(r(\varrho)|_{\varrho\in\mathbb{R}})}{d\varrho}\bigg|_{\varrho=0}-\frac{d(r(\varrho)|_{\varrho\in i\mathbb{R}})}{d\varrho}\bigg|_{\varrho=0}\right)\cdot\exp\left\{4i\sum_{l=1}^{n-1}\gamma_l\right\},$$
$$\kappa^-(\lambda_0;t):=4\lambda_0^4t+\nu(\lambda_0)\ln|t|+\theta^-(\lambda_0)-2\vartheta^-(\lambda_0),$$

and $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

LEMMA 6.3.2. *As* $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$ *and* $(x,t)\in\Omega_n$,

$$((\Psi^{-1}(x,t;0))_{11})^2=\exp\{2\ln(\widetilde{\chi}^c_{22}(0))\}\exp\left\{\frac{2i}{\pi}\int_{\lambda_0}^{\infty}\frac{\ln(1-|r(\varrho)|^2)}{\varrho}d\varrho\right\}\exp\left\{-4i\sum_{l=1}^{n-1}\gamma_l\right\}$$
$$\times\exp\left\{2\ln\left(1-\sum_{i=n}^{N}\left(\frac{2b_i^-}{\lambda_i}+\frac{2\widehat{\delta}_i^-}{\overline{\lambda_i}}\right)\right)\right\}+\mathcal{O}(C(\lambda_0)\exp\{\widetilde{a}\widetilde{b}t\}),$$

where $(\widetilde{\chi}^c_{22}(0))^2$ *is given in Proposition* 6.3.2, $b_i^-=-\frac{\widetilde{\chi}^c_{12}(\lambda_i)}{\widetilde{\chi}^c_{22}(\lambda_i)}\beta_i^-$, $n\le i\le N$, $\beta_j^-,\widehat{\delta}_j^-\sim\mathcal{O}(\exp\{\widetilde{a}^\flat\min_{n+1\le j\le N}|\xi_n-\xi_j|t\})$, $n+1\le j\le N$,

$$\beta_n^-=\frac{\beta_n^{\mathcal{N},-}}{(\widehat{\mathcal{E}}_{nn}^-\widehat{\mathcal{H}}_{nn}^--\widehat{\mathcal{F}}_{nn}^-\widehat{\mathcal{G}}_{nn}^-)},\qquad\widehat{\delta}_n^-=\frac{\widehat{\delta}_n^{\mathcal{N},-}}{(\widehat{\mathcal{E}}_{nn}^-\widehat{\mathcal{H}}_{nn}^--\widehat{\mathcal{F}}_{nn}^-\widehat{\mathcal{G}}_{nn}^-)},$$

with

$$\beta_n^{\mathcal{N},-}:=\frac{g_n^-\widetilde{\chi}^c_{22}(\lambda_n)}{\widetilde{\chi}^c_{11}(\overline{\lambda_n})}+\frac{|g_n^-|^2\widetilde{\chi}^c_{21}(\overline{\lambda_n})\widetilde{\chi}^c_{22}(\lambda_n)}{\overline{\lambda_n}}-\frac{|g_n^-|^2\widetilde{\chi}^c_{22}(\lambda_n)W(\widetilde{\chi}^c_{11}(\overline{\lambda_n}),\widetilde{\chi}^c_{21}(\overline{\lambda_n}))}{\widetilde{\chi}^c_{11}(\overline{\lambda_n})}$$
$$+\frac{2\overline{\lambda_n}|g_n^-|^2\widetilde{\chi}^c_{22}(\lambda_n)\widetilde{\chi}^c_{21}(\overline{\lambda_n})}{(\lambda_n^2-\overline{\lambda_n}^2)},$$

$$\widehat{\mathcal{E}}_{nn}^-\widehat{\mathcal{H}}_{nn}^--\widehat{\mathcal{F}}_{nn}^-\widehat{\mathcal{G}}_{nn}^-:=\frac{1}{\widetilde{\chi}^c_{22}(\lambda_n)\widetilde{\chi}^c_{11}(\overline{\lambda_n})}$$
$$+\frac{\overline{g_n^-}W(\widetilde{\chi}^c_{21}(\overline{\lambda_n}),\widetilde{\chi}^c_{11}(\overline{\lambda_n}))}{\widetilde{\chi}^c_{22}(\lambda_n)\widetilde{\chi}^c_{11}(\overline{\lambda_n})}+\frac{g_n^-W(\widetilde{\chi}^c_{12}(\lambda_n),\widetilde{\chi}^c_{22}(\lambda_n))}{\widetilde{\chi}^c_{11}(\overline{\lambda_n})\widetilde{\chi}^c_{22}(\lambda_n)}$$
$$+\frac{\overline{g_n^-}\widetilde{\chi}^c_{21}(\overline{\lambda_n})}{\overline{\lambda_n}\widetilde{\chi}^c_{22}(\lambda_n)}-\frac{g_n^-\widetilde{\chi}^c_{12}(\lambda_n)}{\lambda_n\widetilde{\chi}^c_{11}(\overline{\lambda_n})}+\frac{(2\overline{\lambda_n})^2|g_n^-|^2\widetilde{\chi}^c_{22}(\lambda_n)\widetilde{\chi}^c_{11}(\overline{\lambda_n})}{(\lambda_n^2-\overline{\lambda_n}^2)^2},$$

$$\widehat{\delta}_n^{\mathcal{N},-}:=\frac{\overline{g_n^-}\,\widetilde{\chi}^c_{21}(\overline{\lambda_n})}{\widetilde{\chi}^c_{22}(\lambda_n)}-\frac{2\overline{\lambda_n}|g_n^-|^2\widetilde{\chi}^c_{11}(\overline{\lambda_n})\widetilde{\chi}^c_{22}(\lambda_n)}{(\lambda_n^2-\overline{\lambda_n}^2)},$$

and $C(\lambda_0)\in\mathcal{L}^\infty(\mathbb{R}_{>M};\mathbb{C})$.

COROLLARY 6.3.1. *As* $t\to-\infty$ *and* $x\to+\infty$ *such that* $\lambda_0>M$ *and* $(x,t)\in\Omega_n$,

$$q(x,t)=Q_{\rm as}^-(x,t)\exp\{i\arg q_{\rm as}^-(x,t)\}+\mathcal{O}\left(\frac{C(\lambda_0)(\ln|t|)^2}{t}\right),$$

where $Q_{\rm as}^-(x,t)$ *is given in Theorem* 2.1, (14)–(29), $\arg q_{\rm as}^-(x,t)$ *is given in Theorem* 2.2, (48)–(51), *and* $C(\lambda_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

COROLLARY 6.3.2. *As* $t\to-\infty$ *and* $x\to-\infty$ *such that* $\widehat{\lambda}_0:=\sqrt{\frac{1}{2}(\frac{x}{t}-\frac{1}{s})}>M$, $\frac{x}{t}>\frac{1}{s}$, $s\in\mathbb{R}_{>0}$, *and* $(x,t)\in\widetilde{\Omega}_n$,

$$u(x,t)=v_{\rm as}^-(x,t)w_{\rm as}^-(x,t)+\mathcal{O}\left(\frac{C(\widehat{\lambda}_0)(\ln|t|)^2}{t}\right),$$

where $v_{\rm as}^-(x,t)$ *and* $w_{\rm as}^-(x,t)$ *are given in Theorem* 2.3, (58)–(70), *and* $C(\widehat{\lambda}_0)\in\mathcal{S}(\mathbb{R}_{>M};\mathbb{C})$.

Acknowledgments. The authors are grateful to P. A. Deift for a copy of [33] prior to publication, A. R. Its for informative discussions, and V. B. Matveev for encouragement and support.

REFERENCES

[1] J. R. TAYLOR, ED., *Optical Solitons - Theory and Experiment*, Cambridge Studies in Modern Optics, Vol. 10, Cambridge University Press, Cambridge, 1992.
[2] G. P. AGRAWAL, *Nonlinear Fiber Optics*, 2nd ed., Academic Press, San Diego, CA, 1995.
[3] A. HASEGAWA AND Y. KODAMA, *Solitons in Optical Communications*, Oxford Series in Optical and Imaging Sciences, No. 7, Oxford University Press, Oxford, 1995.
[4] L. D. FADDEEV AND L. A. TAKHTAJAN, *Hamiltonian Methods in the Theory of Solitons*, Springer-Verlag, Berlin, 1987.
[5] M. WADATI, K. KONNO, AND Y. ICHIKAWA, *A generalization of inverse scattering method*, J. Phys. Soc. Japan, 46 (1979), pp. 1965–1966.
[6] H. EICHHORN, *Application of the inverse scattering method to the generalised non-linear Schrödinger equation*, Inverse Problems, 1 (1985), pp. 193–198.
[7] I. CHEREDNIK, *Basic Methods of Soliton Theory*, Advanced Series in Mathematical Physics, Vol. 25, World Scientific, River Edge, NJ, 1996.
[8] M. W. CHBAT, P. R. PRUCNAL, M. N. ISLAM, C. E. SOCCOLICH, AND J. P. GORDON, *Long-range interference effects of soliton reshaping in optical fibers*, J. Opt. Soc. Amer. B, 10 (1993), pp. 1386–1395.
[9] P. C. SCHUUR, *Asymptotic Analysis of Soliton Problems*, Lecture Notes in Math. 1232, Springer-Verlag, Berlin, 1986.
[10] R. F. BIKBAEV, *On the asymptotics as* $t\to\infty$ *of the Cauchy problem solution for the Landau-Lifshitz equation*, Theoret. and Math. Phys., 77 (1988), pp. 1117–1123.
[11] A. S. FOKAS AND A. R. ITS, *Soliton generation for initial-boundary-value problems*, Phys. Rev. Lett., 68 (1992), pp. 3117–3120.
[12] E. YA. KHRUSLOV AND V. P. KOTLYAROV, *Soliton asymptotics of nondecreasing solutions of nonlinear completely integrable evolution equations*, in Spectral Operator Theory and Related Topics, V. A. Marchenko, ed., Advances in Soviet Mathematics, Vol. 19, AMS, Providence, RI, 1994, pp. 129–180.
[13] V. A. VYSLOUKH AND I. V. CHEREDNIK, *Many-soliton components of solutions of nonlinear Schrödinger equation with perturbing term*, Theoret. and Math. Phys., 78 (1989), pp. 24–31.
[14] Z. CHEN AND N. HUANG, *Explicit N-soliton solution of the modified nonlinear Schrödinger equation*, Phys. Rev. A, 41 (1990), pp. 4066–4069.
[15] A. V. KITAEV AND A. H. VARTANIAN, *Leading-order temporal asymptotics of the modified nonlinear Schrödinger equation: Solitonless sector*, Inverse Problems, 13 (1997), pp. 1311–1339.

[16] M. J. Ablowitz and H. Segur, *Solitons and the Inverse Scattering Transform*, SIAM, Philadelphia, 1981.
[17] M. J. Ablowitz and P. A. Clarkson, *Solitons, Nonlinear Evolution Equations and Inverse Scattering*, London Math. Soc. Lecture Note Series, No. 149, Cambridge University Press, Cambridge, 1991.
[18] S. V. Manakov, *Nonlinear Fraunhoffer diffraction*, Soviet Physics JETP, 38 (1974), pp. 693–696.
[19] V. E. Zakharov and S. V. Manakov, *Asymptotic behavior of non-linear wave systems integrated by the inverse scattering method*, Soviet Physics JETP, 44 (1976), pp. 106–112.
[20] A. R. Its, *Asymptotics of solutions of the nonlinear Schrödinger equation and isomonodromic deformations of systems of linear differential equations*, Soviet Math. Dokl., 24 (1981), pp. 452–456.
[21] S. P. Novikov, S. V. Manakov, L. P. Pitaevskii, and V. E. Zakharov, *Theory of Solitons: The Inverse Scattering Method*, Plenum Press, New York, 1984.
[22] V. E. Zakharov and A. B. Shabat, *Integration of the nonlinear equations of mathematical physics by the method of the inverse scattering transform.* II, Funct. Anal. Appl., 13 (1979), pp. 166–174.
[23] R. Beals, P. Deift, and C. Tomei, *Direct and Inverse Scattering on the Line*, Mathematical Surveys and Monographs, No. 28, AMS, Providence, RI, 1988.
[24] R. Beals and R. R. Coifman, *Scattering and inverse scattering for first order systems*, Comm. Pure Appl. Math., 37 (1984), pp. 39–90.
[25] R. Beals and R. R. Coifman, *Inverse scattering and evolution equations*, Comm. Pure Appl. Math., 38 (1985), pp. 29–42.
[26] A. V. Kitaev, *Self-similar solutions of the modified nonlinear Schrödinger equation*, Theoret. and Math. Phys., 64 (1985), pp. 878–894.
[27] R. F. Bikbaev, *private communication*.
[28] D. J. Kaup and A. C. Newell, *An exact solution for a derivative nonlinear Schrödinger equation*, J. Math. Phys., 19 (1978), pp. 798–801.
[29] V. S. Gerdzhikov, M. I. Ivanov, and P. P. Kulish, *Quadratic bundle and nonlinear equations*, Theoret. and Math. Phys., 44 (1980), pp. 784–795.
[30] J. Lee, *Global solvability of the derivative nonlinear Schrödinger equation*, Trans. Amer. Math. Soc., 314 (1989), pp. 107–118.
[31] X. Zhou, *The Riemann–Hilbert problem and inverse scattering*, SIAM J. Math. Anal., 20 (1989), pp. 966–986.
[32] I. S. Gradshteyn and I. M. Ryzhik, *Tables of Integrals, Series, and Products*, 5th ed., A. Jeffrey, ed., Academic Press, San Diego, CA, 1994.
[33] P. Deift, S. Kamvissis, T. Kriecherbauer, and X. Zhou, *The Toda rarefaction problem*, Comm. Pure Appl. Math., 49 (1996), pp. 35–83.
[34] P. A. Deift and X. Zhou, *Long-time asymptotics for integrable systems. Higher order theory*, Comm. Math. Phys., 165 (1994), pp. 175–191.
[35] A. H. Vartanian, *Higher order asymptotics of the modified nonlinear Schrödinger equation*, Comm. PDE, to appear.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 833–847

EXISTENCE AND BEHAVIOR OF SOLUTIONS TO THE LANDAU–LIFSHITZ EQUATION*

JIAN ZHAI†

Abstract. We prove the existence of nontrivial stable solutions to the Landau–Lifshitz equation with a Neumann boundary condition. The Landau–Lifshitz equation is a phenomenological model for ferromagnets.

Key words. Landau–Lifshitz equation, harmonic map, stable nontrivial solution

AMS subject classifications. 35B35, 35J65, 35K55, 35Q60

PII. S0036141097327951

1. Introduction.

Let Ω be an nonsimply connected bounded domain in $\mathbb{R}^3$ with C^3 boundary. We seek for a solution $u = (u_1, u_2, u_3) : \Omega \to S^2 \subset \mathbb{R}^3$ of the Landau–Lifshitz equation

$$\begin{cases} \Delta u + |\nabla u|^2 u - \lambda(W_u(u) - (W_u(u) \cdot u)u) = 0 \quad \text{in} \quad \Omega, \\ \partial u/\partial \nu = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where $\lambda > 0$ is a parameter and ν is the unit outer normal vector of $\partial\Omega$. In this paper, we assume $W(u) = u_3^2$ and denote $W_u(u) = (0, 0, 2u_3)^t$.

Equation (1.1) is the Euler–Lagrange equation of the Landau–Lifshitz energy functional

$$E_\lambda(u) = \int_\Omega \frac{1}{2}|\nabla u|^2 + \lambda W(u) dx \tag{1.2}$$

on $H^1(\Omega, S^2)$.

Functional (1.2) was first derived for ferromagnetic problems by Landau and Lifshitz [LL] in 1935. The ferromagnetic theory states that below a critical temperature, a sufficiently large ferromagnetic body breaks up into small uniformly magnetized regions separated by thin transition layers. Equation (1.1) is the static equivalent of the time-dependent Landau–Lifshitz equation

$$\begin{aligned} \frac{\partial u}{\partial t} &= -u \times (u \times (\Delta u - \lambda W_u(u))) \\ &\quad + \gamma u \times (\Delta u - \lambda W_u(u)). \end{aligned} \tag{1.3}$$

Here $\gamma \geqq 0$ is called Gilbert damping constant.

It is clear that (1.1) is related to harmonic maps from Ω to S^2. The solutions to (1.1) have similar properties as those of the Ginzburg–Landau equation: there are vortices motions and so on (c.f. [PZ]). Similar properties for the Ginzburg–Landau equation have been discussed recently in a large number of papers (see [BBH], [JMZ],

*Received by the editors September 29, 1997; accepted for publication (in revised form) August 3, 1998; published electronically May 13, 1999. This research was partly supported by the Japan Society for the Promotion of Science.

http://www.siam.org/journals/sima/30-4/32795.html

†Department of Mathematics, Hokkaido University, Sapporo, Japan (zhai@math.sci.hokudai.ac.jp).

etc. and references therein). These results motivate us to conjecture similar conclusions for the Landau–Lifshitz equation.

In this paper, we study the static Landau–Lifshitz equation (1.1) in an nonsimply connected bounded domain of $\mathbb{R}^3$. The existence of nontrivial solutions and their stability are obtained. We shall use the methods developed for the Ginzburg–Landau equation. But the Landau–Lifshitz equation is more complicated: we have to deal carefully with the constraint condition $u \in S^2$, to analyze the spectrum of the linearized operator in a detailed way by using Kato's perturbation theory (c.f. [K]) since the linearized operator is not self-adjoint, and to use some new techniques developed recently for fully nonlinear parabolic equations (c.f. [L]).

Here, we assume $W(u) = u_3^2$, which corresponds to the presence of a continuum of directions of easy magnetization in $S^2 \cap \{u_3 = 0\}$. Our results assert that in each homotopy class from a nonsimply connected bounded domain Ω to $S^2 \cap \{u_3 = 0\}$, there exists a stable distribution of directions of easy magnetization.

This paper consists of five sections and an appendix. In section 2, we state the main theorems of this paper which are proved in sections 3–5. In the Appendix, we modify a stability result given in [H, chapter 5, exercise 6] by the theory of fully nonlinear parabolic equations (see [L], chapter 9) such that it can be applied to the Landau–Lifshitz equation.

2. Main theorems.

THEOREM 1. *Assume that Ω is not simply connected. Then there is $\lambda_0 > 0$ such that for $\lambda > \lambda_0$, there exists a solution,*

$$u_\lambda(x) = (\cos\xi_\lambda(x)\cos\theta_\lambda(x), \cos\xi_\lambda(x)\sin\theta_\lambda(x), \sin\xi_\lambda(x)) \in C^{2+\alpha}(\overline{\Omega}),$$

$0 < \alpha < 1$ to (1.1) corresponding to each homotopy class $[\theta_0]$ of continuous maps from Ω to $S^2 \cap \{u_3 = 0\}$. Moreover, θ_λ is homotopic to θ_0 and

$$\|\xi_\lambda\|_{C^\alpha(\overline{\Omega})} \leqq \frac{C}{\lambda}, \tag{2.1}$$

where the constant C is independent of λ and θ_λ is S^1-valued.

THEOREM 2. *Assume that Ω is not simply connected. Then there exists a $\bar{\lambda} > 0$, and for $\lambda \geqq \bar{\lambda}$, there exists a $\bar{\gamma} = \bar{\gamma}(\lambda) > 0$ such that, for $\lambda \geqq \bar{\lambda}$ and $\gamma \in [0, \bar{\gamma}]$, the solutions $u_\lambda(x)$ obtained in Theorem 1 are stable steady state solutions of the time-dependent Landau–Lifshitz equation*

$$\begin{cases} \dfrac{\partial u}{\partial t} = -u \times (u \times (\Delta u - \lambda W_u(u))) \\ \qquad\qquad + \gamma u \times (\Delta u - \lambda W_u(u)) \quad \textit{in} \quad \Omega \times (0, \infty), \\ \dfrac{\partial u}{\partial \nu} = 0 \quad \textit{on} \quad \partial\Omega \times (0, \infty), \\ u \in S^2 \quad \textit{in} \quad \Omega \times (0, \infty). \end{cases} \tag{2.2}$$

Remark 3. Here stability means Liapunov stability.

Remark 4. The solutions obtained in Theorem 1 take their values inside a neighborhood of the equator of S^2, and thus they can be viewed as maps into a set which is homotopically equivalent to the equatorial S^1. In this sense, we say that the solutions are nontrivial.

3. Proof of Theorem 1. Let $u = (\cos\xi\cos\theta, \cos\xi\sin\theta, \sin\xi)$ $(-\frac{\pi}{2} \leqq \xi \leqq \frac{\pi}{2},\ \theta \in S^1 = \mathbb{R}/2\pi\mathbb{Z})$ in (1.2). The energy functional E_λ is translated into

$$E_\lambda(\theta,\xi) = \int_\Omega \left(\frac{1}{2}|\nabla\xi|^2 + \frac{\cos^2\xi}{2}|\nabla\theta|^2 + \lambda\sin^2\xi\right) dx. \tag{3.1}$$

For any given smooth map $\theta_0 : \overline{\Omega} \to S^1$, $(\theta - \theta_0, \xi)$ can be regarded as a $\mathbb{R}^2$-valued function for any $\theta \in [\theta_0]$ and the Euler–Lagrange equation of (3.1) can be written as

$$\begin{cases} \Delta\xi - \left(\lambda - \dfrac{|\nabla\theta|^2}{2}\right)\sin 2\xi = 0 \quad \text{in} \quad \Omega, \\ \dfrac{\partial\xi}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega \end{cases} \tag{3.2}$$

and

$$\begin{cases} \operatorname{div}(\cos^2\xi\nabla\theta) = 0 \quad \text{in} \quad \Omega, \\ \dfrac{\partial\theta}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{3.3}$$

We first consider the limit functional of (3.1) as $\lambda \to \infty$,

$$E_\infty(\theta) = \int_\Omega \frac{1}{2}|\nabla\theta|^2 dx.$$

Its critical points are well-known harmonic maps to S^1 which satisfy

$$\begin{cases} \Delta\theta = 0 \quad \text{in} \quad \Omega, \\ \dfrac{\partial\theta}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{3.4}$$

The following lemma is standard (the proof can been found, for example, in [JMZ]).

LEMMA 3.1. *Assume that Ω is not simply connected. Then there exists a solution $\theta_* \in C^{2+\alpha}(\overline{\Omega}, S^1)$ to (3.4) in each homotopy class of continuous mappings from $\overline{\Omega}$ into S^1. Moreover, the solution is unique up to an additive constant.*

Hereafter, for technical reasons we fix a point $p \in \Omega$ and let $q = \theta_*(p)$.

DEFINITION 3.2. *Let $\alpha_0 \in (0,1)$ and define*

$$E(\theta_*) = \{\theta | \theta - \theta_* \in C^{1+\alpha_0}(\overline{\Omega}, \mathbb{R}),\ \theta(p) = q,\ \theta \quad \text{is homotopic to} \quad \theta_*, \\ \|\theta - \theta_*\|_{C^{1+\alpha_0}(\overline{\Omega})} \leqq 1\}.$$

LEMMA 3.3. *For any given $\theta \in E(\theta_*)$, there exists a solution ξ_λ^θ to (3.2) which satisfies*

$$\|\xi_\lambda^\theta\|_{C^{2+\alpha_0}(\overline{\Omega})} \leqq C_1, \tag{3.5}$$

$$\lim_{\lambda\to\infty} \sup_{\theta\in E(\theta_*)} \|\xi_\lambda^\theta\|_{C^2(\overline{\Omega})} = 0, \tag{3.6}$$

and

$$\|\xi_\lambda^\theta\|_{C^{\alpha_0}(\overline{\Omega})} \leqq \frac{C_2}{\lambda}, \tag{3.7}$$

provided λ is large enough. Here the constants $C_i = C_i(\|\theta\|_{C^{1+\alpha_0}(\overline{\Omega})})$ $(i = 1, 2)$ are independent of λ.

Proof. Let $\eta := \xi + \frac{C}{\lambda}$, where C is a constant, to be determined in the proof. The equation for η is written as

$$\begin{cases} \Delta\eta - \Big(\lambda - \dfrac{|\nabla\theta|^2}{2}\Big)\sin 2\Big(\eta - \dfrac{C}{\lambda}\Big) = 0 \quad \text{in} \quad \Omega, \\ \dfrac{\partial\eta}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{3.8}$$

Let

$$F(\eta) = -\left(\lambda - \frac{|\nabla\theta|^2}{2}\right)\sin 2\left(\eta - \frac{C}{\lambda}\right).$$

It is easy to check that there exists a constant $C = C(\|\theta\|_{C^{1+\alpha_0}(\overline{\Omega})})$ and $\lambda_0(> 0)$ such that

$$F(0) \geqq 0, \qquad F\Big(\frac{2C}{\lambda}\Big) \leqq 0$$

for $\lambda \geqq \lambda_0$. From [A], there exists a nonnegative solution η_λ^θ to (3.8) which satisfies

$$0 \leqq \eta_\lambda^\theta \leqq \frac{2C}{\lambda},$$

provided $\lambda \geqq \lambda_0$. Let $\xi_\lambda^\theta = \eta_\lambda^\theta - \frac{C}{\lambda}$. Thus ξ_λ^θ is a solution of (3.2) and

$$-\frac{C}{\lambda} \leqq \xi_\lambda^\theta \leqq \frac{C}{\lambda}, \tag{3.9}$$

provided $\lambda \geqq \lambda_0$.

Rewrite (3.2) as

$$-\Delta\xi + 2\lambda\xi = \lambda(2\xi - \sin 2\xi) + \frac{|\nabla\theta|^2}{2}\sin 2\xi,$$

and use the Campanato inequality [C2] to obtain

$$\|\xi\|_{C^{\alpha_0}(\overline{\Omega})} \leqq \frac{C}{\lambda}(\lambda\|2\xi - \sin 2\xi\|_{C^{\alpha_0}(\overline{\Omega})} + \|\xi\|_{C^{\alpha_0}(\overline{\Omega})} + \|\theta\|^2_{C^{1+\alpha_0}(\overline{\Omega})}) \tag{3.10}$$

for a solution $\xi \in C^{\alpha_0}(\overline{\Omega})$ to (3.2).

Claim 1. For $\delta > 0$, there exists $\lambda(\delta) > 0$ such that

$$\|2\xi_\lambda^\theta - \sin 2\xi_\lambda^\theta\|_{C^{\alpha_0}(\overline{\Omega})} \leqq \delta\|\xi_\lambda^\theta\|_{C^{\alpha_0}(\overline{\Omega})},$$

provided $\lambda \geqq \lambda(\delta)$.

Using (3.10) and Claim 1, we get (3.7).

From (3.7) and standard linear elliptic equation theory, we can prove (3.5). Equation (3.6) is a simple corollary of (3.5) and (3.7). □

Proof of Claim 1. For simplicity, we denote $2\xi_\lambda^\theta$ by ξ. Since

$$\sin\xi = \xi - \frac{\xi^3}{3!} + \frac{\xi^5}{5!} - \cdots$$

for small $|\xi|$, from the definition of Hölder seminorm, we have

$$\|\xi - \sin\xi\|_{C^{\alpha_0}(\overline{\Omega})} = \sup_{x,y\in\overline{\Omega},\ x\neq y} \frac{|\xi(x) - \sin\xi(x) - \xi(y) + \sin\xi(y)|}{|x-y|^{\alpha_0}}$$
$$\leqq \|\xi\|_{C^{\alpha_0}(\overline{\Omega})} \sum_{\frac{m-1}{2}=1}^{\infty} \frac{(m-1)(\frac{C}{\lambda})^{m-1}}{m!}.$$

Claim 1 is proved. □

LEMMA 3.4. *For ξ_λ^θ obtained in Lemma* 3.3, *there exists a solution $\theta(\xi_\lambda^\theta)$ to* (3.3) *which satisfies*

$$\|\theta(\xi_\lambda^\theta) - \theta_*\|_{C^{2+\alpha_0}(\overline{\Omega})} \to 0 \quad as \quad \lambda\to\infty \tag{3.11}$$

uniformly for $\theta \in E(\theta_*)$.

Proof. The proof of existence part is similar to Lemma 3.1. So we need only to prove (3.11). For simplicity, we denote ξ_λ^θ and $\theta(\xi_\lambda^\theta)$ by ξ and $\bar\theta$, respectively. From (3.3)–(3.4), we obtain equations for $\bar\theta - \theta_*$:

$$\begin{cases} \operatorname{div}(\cos^2\xi\nabla(\bar\theta-\theta_*)) = -\nabla\cos^2\xi\cdot\nabla\theta_* + (1-\cos^2\xi)\Delta\theta_* \quad \text{in} \quad \Omega, \\ \dfrac{\partial(\bar\theta-\theta_*)}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{3.12}$$

By standard elliptic regularity theory and Lemma 3.3, we have

$$\|\bar\theta - \theta_*\|_{C^{2+\alpha_0}(\overline{\Omega})} \to 0 \quad \text{as} \quad \lambda\to\infty$$

uniformly for $\theta \in E(\theta_*)$. □

PROPOSITION 3.5. *Let $T_\lambda(\theta) = \theta(\xi_\lambda^\theta)$ for $\theta \in E(\theta_*)$. Then $T_\lambda(E(\theta_*))$ is precompact in $E(\theta_*)$ and T_λ is continuous provided λ is large enough.*

Proof. From Lemmas 3.3–3.4, we obtain that $T_\lambda(E(\theta_*))$ is precompact. Let θ_1 and θ_2 be two elements in $E(\theta_*)$. Assume that $\xi_1 = \xi_\lambda^{\theta_1}$, $\xi_2 = \xi_\lambda^{\theta_2}$, and $\bar\theta_1 = \theta(\xi_\lambda^{\theta_1})$, $\bar\theta_2 = \theta(\xi_\lambda^{\theta_2})$ are the solutions of (3.2) and (3.3), respectively. We consider the equations

$$\Delta(\xi_1-\xi_2) - \lambda(\sin 2\xi_1 - \sin 2\xi_2) + \frac{|\nabla\theta_1|^2}{2}\sin 2\xi_1 - \frac{|\nabla\theta_2|^2}{2}\sin 2\xi_2 = 0,$$
$$\operatorname{div}(\cos^2\xi_1\nabla\bar\theta_1) - \operatorname{div}(\cos^2\xi_2\nabla\bar\theta_2) = 0.$$

As in the proof of Lemmas 3.3–3.4, we can prove the continuity of T_λ. □

Proof of Theorem 1. From Proposition 3.5 and the Schauder fixed point theorem, T_λ has a fixed point θ_λ in $E(\theta_*)$ for large λ. Then we obtain a solution $(\theta_\lambda, \xi_\lambda)$ to (3.2)–(3.3) which has the properties stated in Theorem 1. □

4. Proof of Theorem 2 for $\gamma = 0$. By the change of coordinates $u = (\cos\xi\cos\theta, \cos\xi\sin\theta, \sin\xi)$, (2.2) is rewritten as

$$\begin{cases} \partial_t\theta = \left(\dfrac{1}{\cos^2\xi}\right)\operatorname{div}(\cos^2\xi\nabla\theta) \quad \text{in} \quad \Omega\times(0,\infty), \\ \partial_t\xi = \Delta\xi + \left(\dfrac{|\nabla\theta|^2}{2} - \lambda\right)\sin 2\xi \quad \text{in} \quad \Omega\times(0,\infty), \\ \dfrac{\partial\theta}{\partial\nu} = 0, \quad \dfrac{\partial\xi}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega\times(0,\infty). \end{cases} \tag{4.1}$$

For the solution $(\theta_\lambda, \xi_\lambda)$ obtained in Theorem 1, the linearized operator A_λ of the terms in the right-hand side of (4.1) at $(\theta_\lambda, \xi_\lambda)$ can be written as

$$\begin{pmatrix} \Delta + (\frac{1}{\cos^2 \xi_\lambda})\nabla \cos^2 \xi_\lambda \cdot \nabla & -\frac{2}{\cos^2 \xi_\lambda}\nabla\xi_\lambda \cdot \nabla\theta_\lambda - \frac{2\sin \xi_\lambda}{\cos \xi_\lambda}\nabla\theta_\lambda \cdot \nabla \\ (\sin 2\xi_\lambda)\nabla\theta_\lambda \cdot \nabla & \Delta + (|\nabla\theta_\lambda|^2 - 2\lambda)\cos 2\xi_\lambda \end{pmatrix}.$$

Noting that A_λ is not self-adjoint, we decompose A_λ into two parts: $A_\lambda = \bar{A}_\lambda + G$, where

$$\bar{A}_\lambda = \begin{pmatrix} \Delta & -\frac{2}{\cos^2 \xi_\lambda}\nabla\xi_\lambda \cdot \nabla\theta_\lambda \\ 0 & \Delta + (|\nabla\theta_\lambda|^2 - 2\lambda)\cos 2\xi_\lambda \end{pmatrix},$$

and

$$G = \begin{pmatrix} \frac{1}{cos^2\xi_\lambda}\nabla \cos^2 \xi_\lambda \cdot \nabla & -\frac{\sin 2\xi_\lambda}{\cos^2 \xi_\lambda}\nabla\theta_\lambda \cdot \nabla \\ \sin 2\xi_\lambda \nabla\theta_\lambda \cdot \nabla & 0 \end{pmatrix}.$$

Denote $H = L^2(\Omega) \times L^2(\Omega)$ with the standard inner product. It is easy to see that $\bar{A}_\lambda$ can be extended to a self-adjoint operator in H.

We consider the eigenvalue problem for $\bar{A}_\lambda$

$$-\bar{A}_\lambda \begin{pmatrix} \phi \\ \psi \end{pmatrix} = \bar{\mu} \begin{pmatrix} \phi \\ \psi \end{pmatrix}, \tag{4.2}$$

where ϕ and ψ belong to $H^1(\Omega)$ and $\frac{\partial \phi}{\partial \nu} = 0$, $\frac{\partial \psi}{\partial \nu} = 0$ on $\partial\Omega$.

The eigenvalues and eigenfunctions of (4.2) are denoted by

$$\bar{\mu}_1(\lambda) \leqq \bar{\mu}_2(\lambda) \leqq \cdots \leqq \bar{\mu}_k(\lambda) \leqq \cdots$$

and $\{(\phi_k^\lambda, \psi_k^\lambda)\}$, respectively, where $\|\phi_k^\lambda\|^2_{L^2(\Omega)} + \|\psi_k^\lambda\|^2_{L^2(\Omega)} = 1$.

LEMMA 4.1. *For $k \geqq 1$, there exists a constant C_k which is independent of λ such that*

$$\limsup_{\lambda \to \infty} |\bar{\mu}_k(\lambda)| \leqq C_k.$$

Proof. Note that

$$-\int_\Omega (\phi_k^\lambda, \psi_k^\lambda)\bar{A}_\lambda(\phi_k^\lambda, \psi_k^\lambda)^t dx \geqq C(\xi_\lambda, \theta_\lambda) + 2\lambda \int_\Omega (\cos 2\xi_\lambda)(\psi_k^\lambda)^2 dx.$$

Let $\phi_k \in H^1(\Omega)$ ($\|\phi_k\|_{L^2(\Omega)} = 1$) be the kth eigenfunction of Laplace operator $-\Delta$ with the Neumann boundary condition. Then

$$C(\xi_\lambda, \theta_\lambda) \leqq \bar{\mu}_k(\lambda) \leqq -\int_\Omega (\phi_k, 0)\bar{A}_\lambda(\phi_k, 0)^t dx \leqq C(\|\phi_k\|_{H^1(\Omega)}).$$

We get the conclusion of Lemma 4.1. □

From the proof of Lemma 4.1 and (2.1), we also obtain the following lemma.

LEMMA 4.2. *There exists a constant $\bar{C}_k$ such that for large λ,*

$$\left|\lambda \int_\Omega (\psi_k^\lambda)^2 dx\right| \leqq \bar{C}_k.$$

Rewrite the eigenvalue problem (4.2) in the form

$$\begin{cases} \Delta\phi - \dfrac{2}{\cos^2\xi_\lambda}(\nabla\xi_\lambda\cdot\nabla\theta_\lambda)\psi = -\bar{\mu}\phi, \\ \Delta\psi + (|\nabla\theta_\lambda|^2 - 2\lambda)(\cos 2\xi_\lambda)\psi = -\bar{\mu}\psi, \end{cases} \tag{4.3}$$

where ϕ and ψ belong to $H^1(\Omega)$, and $\frac{\partial\phi}{\partial\nu} = 0$, $\frac{\partial\psi}{\partial\nu} = 0$ on $\partial\Omega$.

LEMMA 4.3. *There exists a constant C which is independent of λ such that for large λ,*

$$\|\psi_k^\lambda\|_{C^\alpha(\overline{\Omega})} \leqq \frac{C}{\lambda}, \tag{4.4}$$

$$\|\psi_k^\lambda\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C, \tag{4.5}$$

and

$$\|\phi_k^\lambda\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C. \tag{4.6}$$

Proof. Applying the Campanato inequality to the second equation of (4.3), we obtain

$$\|\psi_k^\lambda\|_{C^\alpha(\overline{\Omega})} \leqq \frac{C}{\lambda}. \tag{4.7}$$

Applying the linear elliptic regularity theory to the second equation again, we get

$$\|\psi_k^\lambda\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C. \tag{4.8}$$

Using (4.8) and the linear elliptic regularity theory for the first equation of (4.3), we obtain that if λ is large enough, then

$$\|\phi_k^\lambda\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C. \tag{4.9}$$

Thus, we proved the lemma. □

By direct calculation, we can prove the following.

LEMMA 4.4. *0 is always the eigenvalue of* (4.3) *and* (4.10). *The corresponding eigenfunction is* $(\frac{1}{\sqrt{|\Omega|}}, 0)$.

$$\begin{cases} -A_\lambda\Phi = \mu\Phi & \text{in} \quad \Omega, \\ \dfrac{\partial\Phi}{\partial\nu} = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{4.10}$$

Next, we consider the spectrum of operator A_λ and analyze its behavior by the perturbation theory (c.f. [K]).

LEMMA 4.5. *Let $T = \rho I - \bar{A}_\lambda$. For $\delta > 0$, there exist $\rho > 0$ and $\lambda_0 = \lambda_0(\delta)$, such that G is T-bounded with T-bound b: $b \leqq \delta^{1/2}$ for $\lambda \geqq \lambda_0$.*

Proof. It is clear that T can be extended to a self-adjoint operator in Hilbert space H and $D(G) \supset D(T)$. For $\Phi = (\phi, \psi)^t \in D(T)$, and for $\delta > 0$, there exists $\lambda_0 = \lambda_0(\delta)$

such that for $\lambda \geqq \lambda_0$ we have

$$\begin{aligned}\|G\Phi\|_H^2 &= \int_\Omega \left(\frac{1}{\cos^2\xi_\lambda}\nabla\cos\xi_\lambda\cdot\nabla\phi - \frac{\sin 2\xi_\lambda}{\cos^2\xi_\lambda}\nabla\theta_\lambda\cdot\nabla\psi\right)^2 + \int_\Omega (\sin 2\xi_\lambda\nabla\theta_\lambda\cdot\nabla\phi)^2\\ &\leqq \max_\Omega\left(2\frac{|\nabla\cos^2\xi_\lambda|^2}{\cos^4\xi_\lambda} + |\sin 2\xi_\lambda\nabla\theta_\lambda|^2\right)\int_\Omega|\nabla\phi|^2\\ &\quad + 2\max_\Omega\left(\frac{|\sin 2\xi_\lambda\nabla\theta_\lambda|}{\cos^2\xi_\lambda}\right)^2\int_\Omega|\nabla\psi|^2\\ &\leqq \delta\langle T\Phi,\Phi\rangle_H\\ &\leqq \delta\|T\Phi\|_H^2\|\Phi\|_H^2,\end{aligned}$$

for some $\rho > 0$. Thus

$$\|G\Phi\|_H \leqq \delta^{1/2}(\|\Phi\|_H + \|T\Phi\|_H). \qquad \square$$

Let $\sigma(T)$ denote the spectral set of T and $\rho(T) = \mathbb{C}\setminus\sigma(T)$. For $\mu\in\sigma(T)$, let

$$d = \operatorname{dist}(\mu, \sigma(T)\setminus\{\mu\})$$

and $\Gamma_r = \{\zeta\in\mathbb{C} : |\zeta-\mu| = r\}$ for $0 < r \leqq d/2$. From [K, Chapter IV, Theorem 3.17], we have following lemma.

LEMMA 4.6. *If* $1+2r+|\mu| < r\delta^{-1/2}$, *then* $\Gamma_r\subset\rho(T-G)$ *and* Γ_r *encloses exactly Multiple*(μ) *eigenvalues of* $T-G$ *and no other points of* $\sigma(T-G)$.

Proof of Theorem 2 *when* $\gamma = 0$. Applying Lemmas 4.1–4.3 to the formula

$$\bar\mu_k(\lambda) = -\int_\Omega(\phi_k^\lambda,\psi_k^\lambda)\bar A_\lambda(\phi_k^\lambda,\psi_k^\lambda)^t dx,$$

we find that as $\lambda\to\infty$,

$$\bar\mu_k(\lambda)\to\mu_k,$$

where μ_k is the kth eigenvalue of the Laplace operator Δ with the Neumann boundary condition. Since

$$0 = \mu_1 < \mu_2 \leqq \mu_3 \leqq \cdots \leqq \mu_k \leqq \cdots,$$

by Lemmas 4.4–4.6, we have

$$0 \equiv \mu_1(\lambda) < \operatorname{Re}(\mu_2(\lambda)),$$

for large λ, where $\mu_k(\lambda)$ denote the kth eigenvalue of the eigenvalue problem (4.10) for A_λ. By [L, chapter 9], we can prove a modification of [H, chapter 5, exercise 6], which is applied to quasi-linear parabolic equations including our case (see the Appendix). From this modification, we get that the solution $(\theta_\lambda,\xi_\lambda)$ is stable provided λ is large enough. Thus we proved Theorem 2 in the case of $\gamma = 0$. $\square$

5. Proof of Theorem 2 when $\gamma \neq 0$. For simplicity, we denote the solution $(\theta_\lambda,\xi_\lambda)$ obtained in section 3 by (θ,ξ). The linearized operator A_λ is written as

$$A_\lambda = \begin{pmatrix} a_{11} & a_{12}\\ a_{21} & a_{22}\end{pmatrix},$$

where

$$a_{11} = \Delta + \frac{\nabla \cos^2 \xi \cdot \nabla}{\cos^2 \xi} - \frac{\gamma \sin 2\xi}{\cos \xi} \nabla\theta \cdot \nabla,$$

$$a_{12} = \begin{pmatrix} -\dfrac{2}{\cos^2 \xi_\lambda} \nabla\xi \cdot \nabla\theta - \dfrac{2 \sin \xi}{\cos \xi} \nabla\theta \cdot \nabla \\ -\dfrac{\gamma}{\cos \xi} (\Delta + (|\nabla\theta|^2 - 2\lambda) \cos 2\xi) \\ -\left(\dfrac{\gamma \sin \xi}{\cos^2 \xi}\right) \left(\Delta\xi + \left(\dfrac{|\nabla\theta|^2}{2} - \lambda\right) \sin 2\xi\right) \end{pmatrix},$$

$$a_{21} = \sin 2\xi \nabla\theta \cdot \nabla + \left(\frac{\gamma}{\cos \xi}\right) (\cos^2 \xi \Delta + \nabla \cos^2 \xi \cdot \nabla),$$

$$a_{22} = \begin{pmatrix} \Delta + (|\nabla\theta|^2 - 2\lambda) \cos 2\xi + \left(\dfrac{\gamma}{\cos \xi}\right) (-\sin 2\xi \Delta\theta \\ -2 \cos 2\xi \nabla\xi \cdot \nabla\theta - \sin 2\xi \nabla\theta \cdot \nabla) + \left(\dfrac{\gamma \sin \xi}{\cos^2 \xi}\right) \operatorname{div}(\cos^2 \xi \nabla\theta) \end{pmatrix}.$$

Decompose the operator A_λ into $\bar{A}_\lambda$ and the perturbation G:

$$\begin{aligned} G &= A_\lambda - \bar{A}_\lambda \\ &= \begin{pmatrix} \frac{\nabla \cos^2 \xi \cdot \nabla}{\cos^2 \xi} - \frac{\gamma \sin 2\xi \nabla\theta \cdot \nabla}{\cos \xi} & \left\{ \begin{array}{c} -\dfrac{\sin 2\xi \nabla\theta \cdot \nabla}{\cos^2 \xi} - \dfrac{\gamma \Delta}{\cos \xi} \\ +\dfrac{2\gamma\lambda}{\cos \xi} (\cos 2\xi + \sin^2 \xi) \end{array} \right\} \\ a_{21} & -\dfrac{\gamma \sin 2\xi \nabla\theta \cdot \nabla}{\cos \xi} \end{pmatrix}. \end{aligned} \tag{5.1}$$

Let $\bar{\mu}_1(\lambda) \leqq \bar{\mu}_2(\lambda) \leqq \cdots \leqq \bar{\mu}_k(\lambda) \leqq \cdots$ and

$$(\bar{\phi}_k^\lambda, \bar{\psi}_k^\lambda), \qquad \|\bar{\phi}_k^\lambda\|^2_{L^2(\Omega)} + \|\bar{\psi}_k^\lambda\|^2_{L^2(\Omega)} = 1, \qquad k = 1, 2, \ldots$$

denote the eigenvalues and eigenfunctions of $\bar{A}_\lambda$, respectively. Then, there exists a constant C which is independent of $\lambda \geqq \bar{\lambda}$ and $\gamma \in [0, \bar{\gamma}]$ for some $\bar{\gamma} > 0$ and $\bar{\lambda} > 0$, such that

$$\begin{aligned} -C + 2\lambda \int_\Omega \cos 2\xi (\bar{\psi}_k^\lambda)^2 dx &\leqq \bar{\mu}_k(\lambda) \\ &= -\int_\Omega (\bar{\phi}_k^\lambda, \bar{\psi}_k^\lambda) \bar{A}_\lambda (\bar{\phi}_k^\lambda, \bar{\psi}_k^\lambda)^t dx \\ &\leqq C, \end{aligned}$$

uniformly for $\lambda \geqq \bar{\lambda}$ and $\gamma \in [0, \bar{\gamma}]$. That is, we have the following lemma.

LEMMA 5.1. *There exist $\bar{\lambda} > 0$, $\bar{\gamma} > 0$, and $C_k > 0$ such that $|\bar{\mu}_k(\lambda)| \leqq C_k$ and*

$$\lambda \int_\Omega (\bar{\psi}_k^\lambda)^2 dx \leqq C_k$$

for $\lambda \geqq \bar{\lambda}$ and $\gamma \in [0, \bar{\gamma}]$, where C_k is independent of λ and γ.

Proof. First we have the estimates

$$\begin{aligned}\bar{\mu}_k(\lambda) &= -\int_\Omega (\bar{\phi}_k^\lambda, \bar{\psi}_k^\lambda)\bar{A}_\lambda(\bar{\phi}_k^\lambda, \bar{\psi}_k^\lambda)^t dx \\ &\geqq \int_\Omega (|\nabla\bar{\phi}_k^\lambda|^2 + |\nabla\bar{\psi}_k^\lambda|^2)dx - C\int_\Omega \gamma(|\nabla\bar{\phi}_k^\lambda \cdot \nabla\bar{\psi}_k^\lambda|)dx - C\int_\Omega (|\nabla\xi\cdot\nabla\bar{\phi}_k^\lambda| \\ &\quad + |\nabla\theta\cdot\nabla\bar{\phi}_k^\lambda| + |\nabla\xi\cdot\nabla\bar{\phi}_k^\lambda| + |\nabla\theta\cdot\nabla\bar{\psi}_k^\lambda|)(|\bar{\phi}_k^\lambda| + |\bar{\psi}_k^\lambda|)dx - C \\ &\quad + 2\lambda\int_\Omega \cos 2\xi(\bar{\psi}_k^\lambda)^2 dx \\ &\geqq \frac{1}{2}\int_\Omega (|\nabla\bar{\phi}_k^\lambda|^2 + |\nabla\bar{\psi}_k^\lambda|^2)dx + 2\lambda\int_\Omega \cos 2\xi(\bar{\psi}_k^\lambda)^2 dx - C \\ &\geqq -C,\end{aligned}$$

provided γ is small enough and λ is large enough, where C only depends on $\|\xi\|_{C^2(\overline{\Omega})}$ and $\|\theta\|_{C^2(\overline{\Omega})}$.

For $\phi \in H^1(\Omega)$ ($\|\phi\|_{L^2(\Omega)} = 1$), we have

$$-\int_\Omega (\phi, 0)\bar{A}_\lambda(\phi, 0)^t dx \leqq C(\|\phi\|^2_{H^1(\Omega)} + 1).$$

Thus we proved the lemma. □

Next the lemma is obtained by direct calculation.

LEMMA 5.2. *0 is always a eigenvalue of operator $\bar{A}_\lambda$ and A_λ. The corresponding eigenfunction is* $(\frac{1}{\sqrt{|\Omega|}}, 0)$.

The eigenvalue problem for $\bar{A}_\lambda$ can be written as

$$\begin{cases} \Delta\phi - \dfrac{2}{\cos^2\xi}(\nabla\xi\cdot\nabla\theta)\psi \\ \qquad -(\dfrac{\gamma}{\cos\xi})(|\nabla\theta|^2(\cos 2\xi)\psi) - \left(\dfrac{\gamma\sin\xi}{\cos^2\xi}\right)\left(\Delta\xi + \dfrac{|\nabla\theta|^2}{2}\sin 2\xi\right)\psi \\ \qquad = -\bar{\mu}\phi \quad \text{in} \quad \Omega, \\ \dfrac{\partial\phi}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega, \end{cases} \tag{5.2}$$

and

$$\begin{cases} \Delta\psi + (|\nabla\theta|^2 - 2\lambda)(\cos 2\xi)\psi + \left(\dfrac{\gamma}{\cos\xi}\right)(-\sin 2\xi\Delta\theta - 2\cos 2\xi\nabla\xi\cdot\nabla\theta)\psi \\ \qquad + \left(\dfrac{\gamma\sin\xi}{\cos^2\xi}\right)\operatorname{div}(\cos^2\xi\nabla\theta)\psi = -\bar{\mu}\psi \quad \text{in} \quad \Omega, \\ \dfrac{\partial\psi}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega. \end{cases} \tag{5.3}$$

LEMMA 5.3. *There exist $\bar{\lambda} > 0$ and $\bar{\gamma} > 0$ such that*

$$\|\bar{\psi}_k^\lambda\|_{C^\alpha(\overline{\Omega})} \leqq \frac{C}{\lambda}, \tag{5.4}$$

$$\|\bar{\psi}_k^\lambda\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C, \tag{5.5}$$

and

$$\|\bar{\phi}_k^{\lambda}\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C, \tag{5.6}$$

provided $\lambda \geqq \bar{\lambda}$ *and* $\gamma \in [0, \bar{\gamma}]$, *where constant* C *only depends on* k *and the* $C^{2+\alpha}(\overline{\Omega})$ *norm of* $(\xi_\lambda, \theta_\lambda)$.

Proof. For simplicity, we denote $(\bar{\phi}_k^{\lambda}, \bar{\psi}_k^{\lambda})$ by (ϕ, ψ).

Step 1. Using the Campanato inequality in (5.3), we find that there exist $\bar{\gamma} > 0$ and $\bar{\lambda} > 0$ such that for $\gamma \in [0, \bar{\gamma}]$ and $\lambda \geqq \bar{\lambda}$,

$$\|\psi\|_{C^{\alpha}(\overline{\Omega})} \leqq \frac{C}{\lambda}, \tag{5.7}$$

where constant C is independent of λ, γ and $\|\phi\|_{C^{2+\alpha}(\overline{\Omega})}$.

By the elliptic regularity theory, there exist $\bar{\gamma} > 0$ and $\bar{\lambda} > 0$ such that for $\gamma \in [0, \bar{\gamma}]$ and $\lambda \geqq \bar{\lambda}$,

$$\|\psi\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C, \tag{5.8}$$

with constant $C < \infty$.

Step 2. Using the boundary extending method, we may extend the Neumann boundary problem (5.2) in Ω to an interior elliptic problem in a larger domain $\tilde{\Omega}$ ($\Omega \subset \tilde{\Omega}$). From the interior boundedness estimate (see, for example, Theorem 8.17 in [GT]) we obtain

$$\sup_{\Omega} |\phi| \leqq C(\|\phi\|_{L^2(\Omega)} + \|\psi\|_{C^{2+\alpha}(\overline{\Omega})} + 1). \tag{5.9}$$

Applying the elliptic regularity theory to (5.2), we have

$$\|\phi\|_{C^{2+\alpha}(\overline{\Omega})} \leqq C(\|\phi\|_{C^{0}(\overline{\Omega})} + \|\psi\|_{C^{2+\alpha}(\overline{\Omega})} + 1). \tag{5.10}$$

By (5.8), (5.9), and (5.10), (5.6) is obtained. □

LEMMA 5.4. *There exist* $\bar{\lambda} > 0$ *and* $d > 0$ *such that for* $\lambda \geqq \bar{\lambda}$,

$$\bar{\mu}_1(\lambda) = 0 \quad \textit{is a simple eigenvalue of} \quad \bar{A}_\lambda,$$

and

$$\bar{\mu}_2(\lambda) \geqq d$$

uniformly for $\gamma \in [0, \bar{\gamma}]$, *where* d *is independent of* λ *and* γ, *and* $\bar{\gamma}$ *is the same as in Lemma* 5.3.

Proof. From Lemma 5.3, (5.2) and (5.3) converge to the eigenvalue problem

$$\begin{cases} \Delta\phi = -\mu\phi & \text{in} \quad \Omega, \\ \dfrac{\partial\phi}{\partial\nu} = 0 & \text{on} \quad \partial\Omega. \end{cases} \tag{5.11}$$

The first eigenvalue of (5.11) is simple and $\mu_1 = 0$. We get the conclusion of this lemma. □

Next, we consider the spectrum of the operator A_λ by a similar method as in section 4. Precisely, we have the following.

LEMMA 5.5. *Let $T = \rho I - \bar{A}_\lambda$. For $\delta > 0$, there exist $\rho > 0$, $\lambda_0 = \lambda_0(\delta)$, and for $\lambda \geqq \lambda_0$ there exists $\gamma_0(\lambda) > 0$ such that for $\lambda \geqq \lambda_0$ and $\gamma \in (0, \gamma_0)$, we have*

$$\|G\Phi\|_H \leqq \delta^{1/2}((\rho+1)\|\Phi\|_H + \|T\Phi\|_H) \quad \textit{for} \quad \Phi \in D(T).$$

That is, G is T-bounded with T-bound $b : b \leqq \delta^{1/2}$.

Proof of Theorem 2. For fixed $\lambda \geqq \bar{\lambda}$, we can choose $\gamma \in (0, \bar{\gamma})$ such that $\gamma\lambda$ is small enough. By a perturbating argument (c.f. [K]), we have that for fixed $\lambda \geqq \bar{\lambda}$, there exists $\bar{\gamma}(\lambda) > 0$ such that for $\gamma \in [0, \bar{\gamma}(\lambda)]$, the eigenvalues $\{\mu_k(\lambda)\}_k$ of operator A_λ have a similar behavior to that of $\bar{A}_\lambda$. Note that 0 is always a eigenvalue of A_λ. Then $\mu_1(\lambda) = 0$ is simple and

$$\mathrm{Re}(\mu_2(\lambda)) \geqq \frac{d}{2}.$$

As in section 4, by Lemma A.1, we proved Theorem 2. □

Appendix. In this appendix, we extend a stability result for a semilinear parabolic equation given in [H, exercise 6, pp. 108] to quasi-linear equations, including the time-dependent Landau–Lifshitz equation. To overcome some technical difficulties, we use the theory for nonlinear parabolic equations (c.f. [L]).

Using spherical coordinates, the time-dependent Landau–Lifshitz equation with the Neumann boundary condition is written as

$$\begin{cases} \eta_t = \dfrac{1}{\cos^2\xi}(\mathrm{div}(\cos^2\xi\nabla(\eta+\theta_0)) + \cos\xi(h_2\cos(\eta+\theta_0) \\ \qquad - h_1\sin(\eta+\theta_0))) - \dfrac{\gamma}{\cos\xi}\left(\Delta\xi + \left(\dfrac{|\nabla(\eta+\theta_0)|^2}{2} - \lambda\right)\sin 2\xi\right. \\ \qquad \left. + H\cdot(-\sin\xi\cos(\eta+\theta_0), -\sin\xi\sin(\eta+\theta_0), \cos\xi)\right) \\ \xi_t = \Delta\xi + \left(\dfrac{|\nabla(\eta+\theta_0)|^2}{2} - \lambda\right)\sin 2\xi \\ \qquad + H\cdot(-\sin\xi\cos(\eta+\theta_0), -\sin\xi\sin(\eta+\theta_0), \cos\xi) \\ \qquad + \dfrac{\gamma}{\cos\xi}(\mathrm{div}(\cos^2\xi\nabla(\eta+\theta_0)) \\ \qquad + \cos\xi(h_2\cos(\eta+\theta_0) - h_1\sin(\eta+\theta_0))) \end{cases} \tag{A.1}$$

in $\Omega \times (0, \infty)$ with the Neumann boundary condition

$$\frac{\partial\eta}{\partial\nu} = 0, \qquad \frac{\partial\xi}{\partial\nu} = 0 \quad \text{on} \quad \partial\Omega \times (0, \infty).$$

Here, $\eta := \theta - \theta_0$, θ_0 is a given map from Ω to $S^2 \cap \{u_3 = 0\}$ with $\partial\theta_0/\partial\nu = 0$ on $\partial\Omega$, and $\theta \in [\theta_0]$. We only need to consider the case of $H \equiv 0$. Denote the right terms by $F(\eta, \xi)$ and (A.1) by

$$y_t = A_\lambda y + \tilde{F}(y); \tag{A.2}$$

here, $y = (\eta, \xi)^T$, A_λ is the linearized operator of F at the steady state solution $(\eta_\lambda, \xi_\lambda) = (\theta_\lambda - \theta_0, \xi_\lambda)$ and

$$\tilde{F}(y) := F(y + (\eta_\lambda, \xi_\lambda)) - A_\lambda y.$$

Let $X = L^2(\Omega, \mathbb{R}^2)$. It is clear that A_λ is a closed sectorial operator from

$$D(A_\lambda) := \{y \in W^{2,2}(\Omega, \mathbb{R}^2) : \partial y/\partial \nu = 0 \text{ on } \partial\Omega\}$$

to X (c.f. [L, pp. 72]). By a direct calculation, we have that $\bar{y}(\tau) = (\tau, 0)^T$ for $\tau \in \mathbb{R}$ satisfies

$$A_\lambda \bar{y}(\tau) + \tilde{F}(\bar{y}(\tau)) = 0.$$

Let $N = \text{span} - \{\bar{y}'(0)\}$ and $X_2 = \overline{R(A_\lambda)}$. Then

$$X = N + X_2.$$

For y given in a neighborhood of 0 in X, we can decompose it into

$$y = \bar{y}(\tau) + z,$$

where z belongs to X_2. Let

$$v \in X : A_\lambda^* v = 0, \langle v, \bar{y}'(0)\rangle_X = 1.$$

Since 0 is a simple eigenvalue of A_λ, (A.2) is equivalent to

$$\text{(A.3)} \qquad \begin{cases} \dfrac{d\tau}{dt} = \phi(\tau, z), \\ \dfrac{dz}{dt} = E_2 A_\lambda z + g(\tau, z), \end{cases}$$

where

$$\phi(\tau, z) = \frac{\langle v, \tilde{F}(\bar{y}(\tau) + z) - \tilde{F}(\bar{y}(\tau))\rangle_X}{\langle v, \bar{y}'(\tau)\rangle_X},$$

E_2 is the projector from X to X_2, and

$$g(\tau, z) = E_2(\tilde{F}(\bar{y}(\tau) + z) - \tilde{F}(\bar{y}(\tau)) - \bar{y}'(\tau)\phi(\tau, z)).$$

LEMMA A.1. *There exists a neighborhood O of 0 in $D := X_2 \cap W^{2,2}(\Omega, \mathbb{R}^2)$ such that $\phi(\tau, \cdot)$ and $g(\tau, \cdot)$ are C^1 functions from O to $\mathbb{R}$ and from O to X, respectively, with a locally Lipschitz continuous derivative, uniformly for $|\tau| \leqq 1$.*

Proof. We have to prove that there exists a neighborhood O of 0 in D, and for $z_0 \in O$ there exist $\delta > 0$ and $C = C(O, \delta) > 0$ such that for $z_1 : \|z_1 - z_0\|_D \leqq \delta$,

$$\|D_z g(\tau, z_0)y - D_z g(\tau, z_1)y\|_X \leqq C\|z_1 - z_0\|_D,$$

$$|D_z \phi(\tau, z_0)y - D_z \phi(\tau, z_1)y| \leqq C\|z_1 - z_0\|_D,$$

uniformly for $y \in D : \|y\|_D = 1$ and $\tau : |\tau| \leqq 1$. Here,

$$\|\cdot\|_D = \|\cdot\|_{W^{2,2}(\Omega)}.$$

By a direct calculation, we find that we only need to prove following estimate:

$$\|(A_\lambda(\bar{y}(\tau) + (\eta_\lambda, \xi_\lambda) + z_0) - A_\lambda(\bar{y}(\tau) + (\eta_\lambda, \xi_\lambda) + z_1))y\|_X \leqq C\|z_0 - z_1\|_D.$$

To obtain the last estimate, we use the Sobolev imbedding inequality. Let $(\eta_i, \xi_i) = \bar{y}(\tau) + (\eta_\lambda, \xi_\lambda) + z_i$ $(i = 0, 1)$. For example, the following inequalities can be proved:

$$\begin{aligned}\int_\Omega \left(\frac{\sin 2\xi_1}{\cos^2 \xi_1}\Delta\eta_1 - \frac{\sin 2\xi_0}{\cos^2 \xi_0}\Delta\eta_0\right)^2 &\leqq C(\xi_1)\int_\Omega (\Delta\eta_1 - \Delta\eta_0)^2 \\ &\quad + \int_\Omega \left(\frac{\sin 2\xi_1}{\cos^2 \xi_1} - \frac{\sin 2\xi_0}{\cos^2 \xi_0}\right)^2 (\Delta\eta_0)^2 \\ &\leqq C(\xi_1)\|\eta_1 - \eta_0\|_D^2 + \left(\int_\Omega (\Delta\eta_0)^2\right)\left\|\frac{\sin 2\xi_1}{\cos^2 \xi_1} - \frac{\sin 2\xi_0}{\cos^2 \xi_0}\right\|_{L^\infty(\Omega)}^2 \\ &\leqq C(\xi_1, \xi_0, \eta_0)(\|\eta_1 - \eta_0\|_D^2 + \|\xi_1 - \xi_0\|_D^2),\end{aligned}$$

and

$$\begin{aligned}\int_\Omega |\nabla(\xi_1 - \xi_0)|^2 |\nabla\eta_0|^2 &\leqq C(\eta_0)\left(\int_\Omega |\nabla(\xi_1 - \xi_0)|^4\right)^{1/2} \\ &\leqq C_1(\eta_0)\|\xi_1 - \xi_0\|_D^2,\end{aligned}$$

etc. Thus, we proved this lemma. □

Next, we apply the theory of [L, chapter 9] to the second equation of (A.3); here we assume that τ satisfies $|\tau| \leqq 1$. It is clear that the graph norm of $E_2 A_\lambda : D(A_\lambda) \cap X_2 \to X_2$ is equivalent to the norm of D. From section 5, we know that 0 is a simple eigenvalue of A_λ for large λ and small γ, and the eigenvalues of $E_2 A_\lambda$ lie in the half plane $\{\omega \in \mathbb{C} : \mathrm{Re}(\omega) \geqq d\}$ for some $d > 0$. Thus we can apply [L, Theorem 9.1.2] to the second equation of (A.3) and obtain that there exists $M > 0$, if $\|z(0)\|_D$ is small; then the solution $z(t)$ satisfies

$$\|z(t)\|_D \leqq Me^{-dt}\|z(0)\|_D \quad \text{for} \quad t \geqq 0. \tag{A.4}$$

Thus, there exists $M' > 0$ such that if $|\tau(0)| + \|z(0)\|_D$ is small, then

$$|\tau(t)| + \|z(t)\|_D \leqq M'e^{-dt}|\tau(0)| + \|z(0)\|_D \quad \text{for} \quad t \geqq 0.$$

That is, the solutions which satisfy the conditions in Theorem 2 are stable in Liapunov meaning.

Acknowledgments. The author is grateful to Prof. Y. Giga and Prof. S. Jimbo for introducing their attention to the Landau–Lifshitz equation and the Ginzburg–Landau equation and for many helpful conversations. The author is also grateful to the referees for many helpful suggestions.

REFERENCES

[A] H. Amann, *On the existence of positive solutions of nonlinear elliptic boundary value problems*, Indiana Univ. Math. J., 21 (1971), pp. 125–146.

[ABV] G. Anzellotti, S. Baldo, and A. Visintin, *Asymptotic behavior of the Landau–Lifshitz model of ferromagnetism*, Appl. Math. Optim., 23 (1991), pp. 171–192.

[BBH] F. Bethuel, H. Brezis, and F. Helein, *Ginzburg-Landau Vortices*, Progress in Nonlinear Differential Equations and Their Applications 13, Birkhäuser Boston, Boston, 1994.

[C1] S. Campanato, *Generation of analytic semigroups by elliptic operators of second order in Hölder spaces*, Ann. Scuola Norm. Sup. Pisa Cl. Sci., 4 (1981), pp. 495–512.

[C2] S. Campanato, *Generation of analytic semigroups in the Hölder topology by elliptic operators of second order with Neumann boundary condition*, Matematiche (Catania), 35 (1980), pp. 61–72.

[GH] B. Guo and M.-C. Hong, *The Landau–Lifshitz equation of the ferromagnetic spin chain and harmonic maps*, Calc. Var. Partial Differential Equations, 1 (1993), pp. 311–334.

[GT] D. Gilbarg and N. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Die Grundlehren der Mathematischen Wissenschaften 224, 2nd ed., Springer-Verlag, New York, 1983.

[H] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics 840, Springer-Verlag, New York, 1981.

[JMZ] S. Jimbo, Y. Morita, and J. Zhai, *Ginzburg-Landau equation and stable solutions in a nontrivial domain*, Comm. Partial Differential Equations, 20 (1995), pp. 2093–2112.

[JZ] S. Jimbo and J. Zhai, *Ginzburg-Landau equation with magnetic effect: Non-simple-connected domains*, J. Math. Soc. Japan, 50 (1998), pp. 663–684.

[K] T. Kato, *Perturbation Theory for Linear Operators*, Die Grundlehren der Mathematischen Wissenschaften 132, Springer-Verlag, New York, 1966.

[L] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Progress in Nonlinear Differential Equations and Their Applications 16, Birkhäuser Verlag, Basel, 1995.

[LL] L.D. Landau and E.M. Lifshitz, *On the theory of the dispersion of magnetic permeability in ferromagnetic bodies*, J. Phys. Z. Sowjetunion, 8 (1935), pp. 153–169; reproduced in Collected Papers of L.D. Landau, Pergamon, New York, 1965, pp. 101–114.

[PZ] N. Papanicolaou and W.J. Zakrzewski, *Dynamics of interacting magnetic vortices in a model Landau–Lifshitz equation*, Phys. D, 80 (1995), pp. 225–245.

[V] A. Visintin, *On Landau–Lifshitz' equations for ferromagnetism*, Japan J. Appl. Math., 2 (1985), pp. 69–84.

[Z] J. Zhai, *Heat flow with tangent penalization converges to mean curvature motion*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998) pp. 875–894.

[Z2] J. Zhai, *Heat flow with tangent penalization*, Nonlinear Anal., 28 (1997), pp. 1333–1346.

[Z3] J. Zhai, *Non-constant stable solutions to Landau–Lifshitz equation*, Calc. Var. Partial Differential Equations, 7 (1998) pp. 159–171.

[Z4] J. Zhai, *Dynamics of domain walls in ferromagnets and weak ferromagnets*, Phys. Lett. A, 234 (1997), pp. 488–492.

[Z5] J. Zhai, *Theoretical velocity of domain wall motion in ferromagnets*, Phys. Lett. A, 242 (1998), pp. 266–270.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 848–878

GENERALIZATION OF THE SCHWARZ REFLECTION PRINCIPLE IN SCATTERING THEORY FOR DISSIPATIVE SYSTEMS: APPLICATION TO PURELY IMAGINARY RESONANT FREQUENCIES*

CHRISTOPHE LABREUCHE†

Abstract. The purpose of this paper is to give an asymptotic estimate of the counting function of the number of resonant frequencies on the purely imaginary axis in the case of the impedance boundary condition. We extend the work of P. Lax, R. Phillips, and J. Beale who studied the Dirichlet, Neumann, and Robin boundary conditions associated with the Helmholtz equation. The method they developed hinges on a special link between the poles of the scattering matrix (i.e., the resonant frequencies) and the zeros of the scattering matrix. This relation holds for conservative boundary conditions but not for absorbing boundary conditions. The first part of this paper consists in finding a relation of this form for the impedance boundary condition. Then we follow the work of P. Lax, R. Phillips, and J. Beale in order to get the final estimate.

Key words. wave equations, scattering theory, integral operators, eigenvalue problem

AMS subject classifications. 35L05, 47A40, 45C05

PII. S0036141097329214

1. Introduction.

In a lot of physical problems, such as radar identification, the analysis of physical measurements is done through the transient response. The behavior of this latter is explained by the existence of complex frequencies called the *"resonant frequencies"* (see the Singularity Expansion Method [1]). In the view of inverse problems, one can wonder what can be deduced of the obstacle from a knowledge of its resonant frequencies. We wish to infer from these numbers an estimate of the size of the obstacle. This is of great practical interest. This is also mathematically very important since this piece of information can remove the ill-posedness of the inverse problem [8].

Let Ω^{int} be an open bounded domain and Ω its complement in $\mathbb{R}^N$. The outward normal to $\Gamma = \partial\Omega^{\text{int}}$ is denoted by $\mathbf{n}$. We assume that Γ is twice differentiable. The resonant frequencies are defined as the poles of the scattering matrix or equivalently the singularities of the *"outgoing"* Green's function for the Helmholtz equation outside an obstacle [9]. A lot of work has been done on the location of the resonant frequencies (see [16] for an overview). There are in fact two classes of results on poles: counting functions of the number of poles and pole-free regions. We focus in this paper only on the first class. Several results dealing with pole-free regions can be found in [11, 13, 18]. It is known that the set of all resonant frequencies has no cluster point, except at infinity. Hence in any compact set of $\mathbb{C}$, there is only a finite number of poles. Let us denote by $N_{\text{G}}(\sigma)$ the number of poles within the circle $|z| < \sigma$. R. Melrose has shown in [15] the following sharp polynomial bound in the case of the Dirichlet boundary condition

$$N_{\text{G}}(\sigma) \leq C_{\text{G}} + C_{\text{G}}\,\sigma^N, \tag{1}$$

*Received by the editors October 27, 1997; accepted for publication July 22, 1998; published elctronically May 13, 1999. This work was carried out while the author was visiting the University of Delaware, Newark, DE.

http://www.siam.org/journals/sima/30-4/32921.html

†Thomson CSF, LCR, Domaine de Corbeville, 91404 Orsay Cedex, France (christophe.labreuche@lcr.thomson-csf.com).

where N is the dimension of space. For this result as well as for most results on poles, the dimension of space is assumed to be odd. Inequality (1) has no application to the inverse problem, since no lower bound of $N_{\mathrm{G}}(\sigma)$ is known and no information about C_{G} is at our disposal.

The second result concerns the counting function on the purely imaginary axis for odd space dimensions:

$$N_{\mathrm{I}}(\sigma) = \#\{k \text{ pole }, \ k \in \mathbb{R} \text{ and } |k| < \sigma\}\,.$$

In odd space dimensions, we know that there are infinitely many poles on the purely imaginary axis. More precisely, we have

$$\frac{1}{(N-1)!}\left(\frac{R_1}{\gamma_0}\right)^{N-1} \leq \lim_{\sigma\to\infty} \frac{N_{\mathrm{I}}(\sigma)}{\sigma^{N-1}} \leq \frac{1}{(N-1)!}\left(\frac{R_2}{\gamma_0}\right)^{N-1}, \tag{2}$$

where R_1 is the radius of the largest sphere contained in Ω^{int}, R_2 is the radius of the smallest sphere containing Ω^{int}, and γ_0 is a known constant. This result has been shown by P. Lax and R. Phillips in [10] for the Dirichlet and the Neumann boundary conditions, and by J. Beale in [2] for the Robin boundary condition. The aim of this work is to show (2) in the case of the impedance boundary condition.

Let us now give the idea of the proof of (2) in the case of the Dirichlet, Neumann, and Robin boundary conditions. These three boundary conditions are very close to one another since they share a very important property: the energy of the acoustic waves in the whole exterior domain is conserved. The conservative nature of the boundary condition means that the scattering matrix $\mathcal{S}(k)$ is unitary for $k \in \mathbb{R}$. When k is not real, there is a more general relation (namely the Schwarz reflection principle):

$$\mathcal{S}(\overline{k}) = [\mathcal{S}^*(k)]^{-1}\,. \tag{3}$$

It follows that $\overline{k}$ is a zero of $\mathcal{S}$ if and only if k is a pole of $\mathcal{S}^*$. On the other hand, the poles of $\mathcal{S}$ are exactly the poles of $\mathcal{S}^*$ (see Lemma 23). Therefore the location of the zeros of $\mathcal{S}$ is linked to the location of the poles of $\mathcal{S}$, as stated in the following rule: k is a pole of $\mathcal{S}$ if and only if $\overline{k}$ is a zero of $\mathcal{S}$. Consequently, the repartitioning of the resonant frequencies on the purely imaginary axis is symmetric to the repartitioning of the zeros of $\mathcal{S}$ on the purely imaginary axis. This property is at the root of (2), since it is easier to study the zeros of an operator than it is to study its poles. A special analysis of the zeros of the scattering matrix can be carried out on a purely imaginary axis since the study of the zeros of $\mathcal{S}(k)$ for $k \in i\mathbb{R}^+$ turns into a real symmetric eigenvalue problem for an elliptic operator. As a matter of fact, the scattering matrix reads

$$\mathcal{S}(k) = (-1)^{\frac{N+1}{2}}\,[\mathrm{I} + Q(k)W]\,, \tag{4}$$

where $Q(k)$ is a compact operator and W is a unitary operator. Moreover, $Q(k)$ is also self-adjoint on the purely imaginary axis. Hence, $N_{\mathrm{I}}(\sigma)$ is the number of times -1 is an eigenvalue of $Q(i\mu)W$ (where $Q(i\mu)$ is compact and self-adjoint) when μ runs in the slab $[0, \sigma]$.

The general theory of the Lax–Phillips group for dissipative systems can be found

in [12], the following problem is studied

$$\left\{\begin{array}{ll} \dfrac{\partial^2 w(t,\mathbf{x})}{\partial t^2} - \Delta w(t,\mathbf{x}) = 0\ , & \text{in} \quad \mathbb{R}^+\times\Omega \\ \dfrac{\partial w(t,\mathbf{x})}{\partial \mathbf{n}} - \lambda(\mathbf{x})\dfrac{\partial w(t,\mathbf{x})}{\partial t} = 0, & \text{on} \quad \mathbb{R}^+\times\Gamma \\ w(0,\mathbf{x}) = f_0(\mathbf{x}) \quad \text{in} \quad \Omega & \\ \dfrac{\partial w(0,\mathbf{x})}{\partial t} = f_1(\mathbf{x}) \quad \text{in} \quad \Omega. & \end{array}\right. \tag{5}$$

The impedance λ is assumed to satisfy the physical hypotheses

$$0 \leq \lambda(\mathbf{x}) < 1, \qquad \forall \mathbf{x} \in \Gamma. \tag{6}$$

The time translation operator $U(t)$ is defined by

$$U(t)\ :\ \left(\begin{array}{c} f_0 \\ f_1 \end{array}\right) \mapsto \left(\begin{array}{c} w(t,.) \\ \frac{\partial w(t,.)}{\partial t} \end{array}\right).$$

The scattering theory developed by P. Lax and R. Phillips in [9] consists of studying the properties of the one-parameter semigroup $\{U(t)\}_{t\in\mathbb{R}}$ acting on $L^2(\Omega)\times L^2(\Omega)$. P. Lax and R. Phillips showed that $U(t)$ has some eigenvalues

$$U(t)g = e^{-ikt}g$$

for some $g \in L^2_{loc}(\Omega)\times L^2_{loc}(\Omega)$ and some discrete values $k \in \mathbb{C}$. The factors k are the resonant frequencies . The first component u of g is a nontrivial solution to

$$\left\{\begin{array}{ll} \Delta u + k^2 u = 0 & \text{in} \quad \Omega, \\ \frac{\partial u}{\partial \mathbf{n}} + ik\lambda u = 0 & \text{on} \quad \Gamma, \\ u \quad \text{outgoing.} & \end{array}\right. \tag{7}$$

In section 2, we will explain what is meant by *"outgoing."*

We wish to extend (2) to the impedance problem described above. Problem (5) corresponds to the fact that the obstacle is absorbing some energy. The main difficulty due to this boundary condition is that the scattering matrix is no longer unitary for real frequencies. Moreover (3) does not hold, which implies that there does not seem to be any link between the zeros and the poles of $\mathcal{S}$. For this reason, as noticed in [2], the proof of (2) seems to be nonachievable.

To generalize the Schwarz reflection principle, we find it more convenient to focus on the problem in frequency (i.e., (7)) rather than the problem in time (i.e., (5)). This enables us to introduce a more general impedance boundary condition, $\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u = g$, where $\zeta(k)$ is the impedance function (which depends on $k \in \mathbb{C}$ and on $\mathbf{x} \in \Gamma$). $\zeta(k)$ is supposed to be analytic in k and to satisfy the following physical assumption:

$$\Re\big(\zeta(k)\,\overline{k}\big) \geq 0 \quad \forall k \in \mathbb{C}. \tag{8}$$

The positive sign means that the obstacle is absorbing some energy. Obviously, this condition is satisfied with the impedance function $\zeta(k) = \lambda k$.

By studying more deeply the form of the scattering matrix, a relation quite similar to (3) can be derived. To allow this, the impedance $\zeta(k)$ must be put as an argument of the scattering matrix: we denote by $\mathcal{S}(k,\zeta(k))$ the scattering matrix at the frequency k with the impedance function $\zeta(k)$. With this notation, we will show that

$$\mathcal{S}\Big(\overline{k}, -\overline{\zeta(k)}\Big) = [\mathcal{S}^*(k,\zeta(k))]^{-1}. \tag{9}$$

In other words, k is a pole of $\mathcal{S}$ with the impedance function $\zeta(k)$ if and only if $\overline{k}$ is a zero of $\mathcal{S}$ with the impedance function $-\overline{\zeta(k)}$. The application of (3) to this case would lead to this equivalence: k is a pole of $\mathcal{S}$ with the impedance function $\zeta(k)$ if and only if $\overline{k}$ is a zero of $\mathcal{S}$ with the impedance function $\zeta(\overline{k})$. Consequently, the use of (9) implies a modification of the impedance function considered when studying the zeros of $\mathcal{S}$. If $(k, \zeta(k))$ satisfies (8), then $\left(\overline{k}, -\overline{\zeta(k)}\right)$ does not satisfy condition (8) since

$$\Re\left(-\overline{\zeta(k)}\,\overline{\overline{k}}\right) = -\Re\big(\zeta(k)\,\overline{k}\big) \leq 0 \quad \forall k \in \mathbb{C}.$$

Therefore, we shall need to define and study the scattering matrix for nonphysical impedances. However, we assume that the nonphysical impedances satisfy the following growth condition

$$\left|\Re\big(\zeta(k,\mathbf{x})\overline{k}\big)\right| \leq \lambda_m |k|^2 \quad \forall k \in \mathbb{C} \text{ with } |k| \geq K \text{ and } \forall \mathbf{x} \in \Gamma, \tag{10}$$

for some $K \geq 0$ and $0 < \lambda_m < 1$. At this point, we are ready to apply the work of P. Lax, R. Phillips, and J. Beale in order to give an estimate of the number $N_{\mathrm{I}}(\sigma)$ of purely imaginary resonant frequencies for the impedance boundary condition $\frac{\partial w}{\partial \mathbf{n}} - \lambda \frac{\partial w}{\partial t} = 0$ whose modulus is lower than σ. From (9), it suffices to count the number of zeros of $\mathcal{S}(i\mu, -i\lambda\mu)$ when μ runs in $[0, \sigma]$.

One can hint at formula (2) from a completely different method than that developed in the rest of this paper.

THEOREM 1. *There exists a constant C such that for σ large enough, we have*

$$N_{\mathrm{I}}(\sigma) \leq C\sigma^{N-1} \ .$$

Proof. The scattering matrix reads (see (45))

$$\mathcal{S}(i\sigma, -i\sigma\lambda) = (-1)^{\frac{N+1}{2}} \left[\mathrm{I} - \left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma) W\right],$$

where $s(\sigma)$ is the so-called *transmission coefficient*. We will show in Lemma 27 that for σ large enough, $s(\sigma)$ is a negative and self-adjoint operator. Consequently, the singular values [6] $s_j(s(\sigma))$ of $s(\sigma)$ are identical to the eigenvalues $\pm\lambda_j(s(\sigma)W)$ of $\pm s(\sigma)W$ (the $\pm$ sign comes from W). From Theorem 34, the number of poles in the slab $[-i\sigma, 0]$ is asymptotically (as $\sigma \to \infty$) equal to the number of eigenvalues of $\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)W$ that are greater than 1. Up to a factor 2, this is also asymptotically equal to the number of singular values of $\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)$ that are greater than 1. From (32) and (33), the kernel of $s(\sigma)$ is

$$-\frac{\sigma^{\frac{N-3}{2}}}{2(2\pi)^{\frac{N-1}{2}}} \int_\Gamma \left(\frac{\partial e^{\sigma \hat{\mathbf{y}}\cdot\mathbf{z}}}{\partial \mathbf{n}_\mathbf{z}} + \lambda\sigma e^{\sigma\hat{\mathbf{y}}\cdot\mathbf{z}}\right) S^+(i\sigma, -i\lambda\sigma) \left(\frac{\partial e^{\sigma\hat{\mathbf{x}}\cdot\mathbf{z}}}{\partial \mathbf{n}_\mathbf{z}} + \lambda\sigma e^{\sigma\hat{\mathbf{x}}\cdot\mathbf{z}}\right) d\gamma(\mathbf{z}),$$

where the jump operator $S^+(i\sigma, -i\lambda\sigma)$ is defined in section 2. Let us denote by Δ_S the Laplace–Beltrami operator on the unit sphere S_1. Using the classical estimate $\left|(\Delta_S + 1)^{m/2} e^{\sigma\hat{\mathbf{x}}\cdot\mathbf{z}}\right| \leq C m^m e^{C\sigma}$ for $\mathbf{z} \in \Gamma$, we get for an arbitrary integer m:

$$\left\| \left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} (\Delta_S + 1)^{m/2}\, s(\sigma) \right\|_{L^2(S_1)\to L^2(S_1)} \leq C m^m e^{C\sigma} \ .$$

Writing $s(\sigma) = (\Delta_S + 1)^{-m/2} (\Delta_S + 1)^{m/2} s(\sigma)$, we have [6]

$$s_j\left(\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)\right)$$
$$\leq s_j\left((\Delta_S+1)^{-m/2}\right)\left\|\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} (\Delta_S+1)^{m/2} s(\sigma)\right\|_{L^2(S_1)\to L^2(S_1)} .$$

Since $(\Delta_S + 1)^{-m/2}$ is pseudodifferential operator of order $-m$, there exists C such that $s_j\left((\Delta_S+1)^{-m/2}\right) \leq \frac{C}{j^{m/(N-1)}}$ (see, for instance, [7, Lemma A.4]). Henceforth

$$s_j\left(\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)\right) \leq C\frac{m^m}{j^{m/(N-1)}} e^{C\sigma} .$$

Taking the value of m which minimizes the right-hand side, we obtain

$$s_j\left(\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)\right) \leq \exp\left(C\sigma - \frac{1}{C} j^{\frac{1}{N-1}}\right).$$

It follows that the condition $s_j\left(\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)\right) > 1$ implies that $j \leq C\sigma^{N-1}$ for some constant C. Thus,

$$\#\left\{j \ , \ s_j\left(\left(\frac{-\sigma}{2\pi}\right)^{\frac{N-1}{2}} s(\sigma)\right) > 1\right\} \leq C\sigma^{N-1} .$$

From the above remarks, this gives the final estimate. □

We would like to point out a difference of notation between this paper and [9]. The resonant frequencies as defined here lie in the lower half plane of complex frequencies (see Theorem 16) whereas the resonant frequencies as defined in [9] lie in the upper half plane. This is just a matter of convenience and habit.

Let us now give some notation. By $\mathcal{L}(X, Y)$ is meant the set of all the linear operators that maps X onto Y continuously. $L^2(\mathcal{O})$ is the classical space of all square integrable distributions. $H^s(\mathcal{O})$ is the classical Sobolev space ($s \in \mathbb{R}$). $H^s_{loc}(\mathcal{O})$ is the Frechet space of the distributions whose restriction to each bounded domain $\mathcal{O}' \subset \mathcal{O}$ is in $H^s(\mathcal{O}')$. $\|.\|_{L^2(\mathcal{O})}$ and $\|.\|_{H^s(\mathcal{O})}$ are the classical norms in $L^2(\mathcal{O})$ and $H^s(\mathcal{O})$, respectively. B_R is the ball of center 0 and radius R. S_R is the sphere of center 0 and radius R. S_1 in the unit sphere. The complex scalar product in $L^2(\Gamma)$ is denoted by $\langle ., . \rangle$, where Γ is the boundary of an obstacle. $(., .)$ stands for the complex scalar product in $L^2(S_1)$. The two Hankel functions of order 0 and first and second kind are denoted by $H_0^{(1)}(z)$ and $H_0^{(2)}(z)$. For a linear operator B applying on complex valued functional spaces such as L^2 or H^1, let us define its conjugate $\overline{B}$ as follows: $\overline{B}g$ is the complex conjugate of $B\overline{g}$. For a function u defined in either Ω^{int} or Ω, we denote by $u_{|\Gamma}$ the restriction of u on Γ. Let us now consider a function v defined on $\Omega^{\text{int}} \cup \Omega = \mathbb{R}^N \backslash \Gamma$. When a function v is considered on Γ, v^{int} or ${v^{\text{int}}}_{|\Gamma}$ denotes the value on Γ of the restriction to Ω^{int}, whereas v or $v_{|\Gamma}$ means that we take the value on Γ of the restriction to Ω. The jump of v on Γ is defined by $[\![v]\!] := {v^{\text{int}}}_{|\Gamma} - v_{|\Gamma}$. In this paper, vectors will be typed in bold characters, whereas scalar numbers will be typed with standard letters. For $\mathbf{x} \in \mathbb{R}^N$, we define $\hat{\mathbf{x}} := \frac{\mathbf{x}}{|\mathbf{x}|} \in S_1$.

The layout of this paper is as follows. In section 2, we study the nonphysical impedance problem (i.e., subject to (10)) and introduce the notion of incoming and outgoing solutions. Then we define the far field pattern of these solutions. We also give a very important relation between the incoming and the outgoing far field patterns. In section 3, we focus on the physical problem (i.e., subject to (8) and (10)). We define the resonant frequencies as the poles of the outgoing Green function. Then the expression of the scattering matrix is given. The resonant frequencies are proved to be exactly the poles of the scattering matrix. We show (9) and state the relation between the poles and the zeros of $\mathcal{S}$. In section 4, inequality (2) is proved (with the impedance $\zeta(k) = \lambda k$).

2. Definition and properties of the far field pattern. In this section, we aim to construct the far field pattern of the incoming and the outgoing solutions and give the link between these two quantities. Here we consider the general impedance boundary condition $\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u = g$ on Γ, where $\zeta(k)$ is an analytic function of $k \in \mathbb{C}$ which satisfies

$$\Re(\zeta(k)) \neq 0 \quad \forall k \in \mathbb{R}^*. \tag{11}$$

This condition means that the energy of the system is not conserved (see Lemma 19). The impedance function ζ depends also on $\mathbf{x} \in \Gamma$, and will be denoted either $\zeta(k)$ or $\zeta(k, \mathbf{x})$. We assume furthermore that $\zeta(k)$ belongs to $C^2(\Gamma)$ and satisfies

$$\left|\Re\big(\zeta(k, \mathbf{x})\overline{k}\big)\right| \leq \lambda_m |k|^2 \quad \forall k \in \mathbb{C} \text{ with } |k| \geq K \text{ and } \forall \mathbf{x} \in \Gamma \tag{12}$$

for some $K \geq 0$ and $0 < \lambda_m < 1$.

2.1. Incoming and outgoing solutions. Let us consider the following problem for $\Im(k) > 0$:

$$\begin{cases} u^+ \in H^1(\Omega)\,, \\ -\Delta u^+ - k^2 u^+ = 0 \quad \text{in} \quad \Omega, \\ \frac{\partial u^+}{\partial \mathbf{n}} + i\zeta(k)u^+ = g \quad \text{on} \quad \Gamma. \end{cases} \tag{13}$$

LEMMA 2. *Let $\zeta(k)$ be an impedance function satisfying condition* (12)*. Then the problem* (13) *has a unique solution when the complex frequency k lies inside the set $\mathcal{C}^+$ defined by*

$$\Im(k) \geq \lambda_m |k| + \lambda_m C + 1 \text{ and } |k| \geq K,$$

where $C > 0$ depends only on the domain Ω (see Figure 1*).*

The set $\mathcal{C}^+$ is not empty since $\lambda_m < 1$. The condition $\lambda_m < 1$ is necessary to show the coerciveness of (13).

Proof. One can give the variational formulation of (13):

$$b^+(u^+, v) = -\int_\Gamma g\overline{v}\, d\gamma \quad \forall v \in H^1(\Omega),$$

where

$$b^+(u, v) := \int_\Omega \big(\nabla u \cdot \nabla \overline{v} - k^2 u\overline{v}\big) - \int_\Gamma i\zeta(k) u\overline{v}\, d\gamma.$$

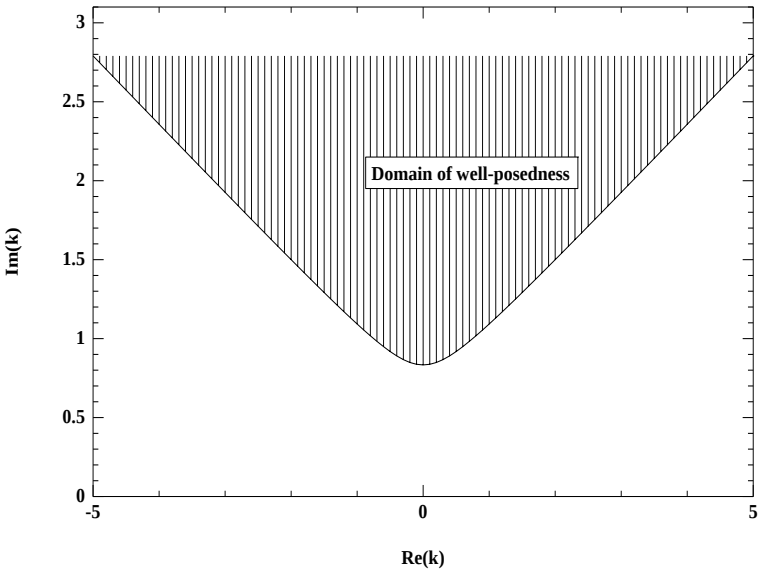


FIG. 1. *The set* $\mathcal{C}^+$.

In order to use Lax–Milgram's lemma, let us compute

$$\Re\big(i\overline{k}b^+(u,u)\big) = \Im(k)\int_\Omega |\nabla u|^2 + \Im(k)|k|^2\int_\Omega |u|^2 + \int_\Gamma \Re\big(\zeta(k)\overline{k}\big)\,|u|^2\,d\gamma. \tag{14}$$

By (12), we have for $|k| \geq K$

$$\int_\Gamma \Re\big(\zeta(k)\overline{k}\big)\,|u|^2\,d\gamma \geq -\lambda_m|k|^2\int_\Gamma |u|^2\,d\gamma.$$

From Lemma 3.3 in [2], for any domain $\mathcal{D}$ with a twice differentiable and bounded boundary, there exists a constant C depending only on $\mathcal{D}$ such that $\forall\ u \in H^1(\mathcal{D})$ and $\forall\ \epsilon > 0$

$$\int_{\partial\mathcal{D}} |u|^2 \leq \epsilon\int_{\mathcal{D}} |\nabla u|^2 + \left(C + \frac{1}{\epsilon}\right)\int_{\mathcal{D}} |u|^2. \tag{15}$$

Then, by (12), using (15) with $\mathcal{D} = \Omega$ and $\epsilon = \frac{1}{|k|}$, we get

$$\begin{aligned}\int_\Gamma \Re\big(\zeta(k)\overline{k}\big)\,|u|^2\,d\gamma &\geq -\lambda_m|k|^2\int_\Gamma |u|^2\,d\gamma \\ &\geq -\lambda_m|k|\|\nabla u\|^2_{L^2(\Omega)} - \lambda_m|k|^2(C+|k|)\|u\|^2_{L^2(\Omega)}\,,\end{aligned}$$

and

$$\Re\big(i\overline{k}b^+(u,u)\big) \geq (\Im(k) - \lambda_m|k|)\|\nabla u\|^2_{L^2(\Omega)} + |k|^2(\Im(k) - \lambda_m|k| - \lambda_m C)\|u\|^2_{L^2(\Omega)}\,.$$

Therefore, if $\Im(k) \geq \lambda_m|k| + \lambda_m C + 1$, then we have shown the coerciveness of $i\overline{k}b^+$:

$$\Re\big(i\overline{k}b^+(u,u)\big) \geq \min\big(\lambda_m C + 1, |k|^2\big)\|u\|^2_{H^1(\Omega)}\,.$$

This means, by Lax–Milgram's lemma, that (13) has a unique solution. The lemma is proved. □

Remark 3. It is well known (see [4]) that when $\Im(k) > 0$, the requirement that the solution of the Helmholtz equation belongs to $H^1(\Omega)$ is equivalent to the outgoing Sommerfeld condition [4]. Namely, whenever $\Im(k) > 0$, (13) is equivalent to the following problem

$$\begin{cases} u^+ \in H^1_{loc}(\Omega) \\ -\Delta u^+ - k^2 u^+ = 0 \quad \text{in} \quad \Omega, \\ \frac{\partial u^+}{\partial \mathbf{n}} + i\zeta(k) u^+ = g \quad \text{on} \quad \Gamma, \\ \lim_{r\to\infty} r^{\frac{N-1}{2}} \left(\frac{\partial u^+}{\partial r} - iku^+ \right) = 0 \quad \text{uniformly in all directions.} \end{cases}$$

This defines the *"outgoing"* solution of the Helmholtz equation.

For $k \in \mathcal{C}^+$ we define the operator $R^+(k, \zeta(k))$ by $R^+(k, \zeta(k))\, g = u^+$, where u^+ is the solution to (13). In order to extend R^+ in $\mathbb{C}\backslash\mathcal{C}^+$, we need a formulation of the solution by integral equations over the boundary Γ (see [5]). u^+ can be written as a combination of the single layer and the double layer potentials

$$u^+ = -i\tilde{V_k^+}\left(\zeta(k) v^+\right) - \tilde{K_k^+}\left(v^+\right) \qquad \text{in} \quad \Omega, \tag{16}$$

where the potential $v^+ \in H^{1/2}(\Gamma)$ defined on Γ must be determined. The integral operators

$$\begin{aligned} \tilde{V_k^+} v(\mathbf{x}) &:= \int_\Gamma G_k^+(\mathbf{x}, \mathbf{y})\, v(\mathbf{y})\ d\gamma(\mathbf{y}), \\ \tilde{K_k^+} v(\mathbf{x}) &:= \int_\Gamma \frac{\partial}{\partial \mathbf{n_y}} G_k^+(\mathbf{x}, \mathbf{y})\, v(\mathbf{y})\ d\gamma(\mathbf{y}) \end{aligned}$$

are defined for $\mathbf{x} \notin \Gamma$. We recall that the outgoing fundamental solution $G_k^+(\mathbf{x}, \mathbf{y})$ is equal to $\frac{i}{4} H_0^{(1)}(k|\mathbf{x} - \mathbf{y}|)$ for $N = 2$ and $\frac{\exp(ik|\mathbf{x}-\mathbf{y}|)}{4\pi|\mathbf{x}-\mathbf{y}|}$ for $N = 3$. The advantage of integral representations is that the Helmholtz equation and the outgoing Sommerfeld condition are automatically satisfied. Then the potential $v^+ \in H^{1/2}(\Gamma)$ is determined by requiring that the impedance boundary condition is satisfied, leading to [5]:

$$\begin{aligned} T^+(k, \zeta(k))\, v^+ := &-D_k^+\left(v^+\right) - i\Big(\zeta(k) K_k^+\left(v^+\right) + \left(K_k^+\right)^t\left(\zeta(k) v^+\right)\Big) \\ &+\zeta(k) V_k^+\left(\zeta(k) v^+\right) = g, \end{aligned} \tag{17}$$

where the integral operators

$$\begin{aligned} V_k^+(v)(\mathbf{x}) &:= \int_\Gamma G_k^+(\mathbf{x}, \mathbf{y})\, v(\mathbf{y})\ d\gamma(\mathbf{y}), \\ K_k^+(v)(\mathbf{x}) &:= \int_\Gamma \frac{\partial G_k^+(\mathbf{x}, \mathbf{y})}{\partial \mathbf{n_y}} v(\mathbf{y})\ d\gamma(\mathbf{y}), \\ \left(K_k^+\right)^t(v)(\mathbf{x}) &:= \int_\Gamma \frac{\partial G_k^+(\mathbf{x}, \mathbf{y})}{\partial \mathbf{n_x}} v(\mathbf{y})\, d\gamma(\mathbf{y}), \\ D_k^+(v)(\mathbf{x}) &:= \int_\Gamma \frac{\partial^2 G_k^+(\mathbf{x}, \mathbf{y})}{\partial \mathbf{n_x} \partial \mathbf{n_y}} v(\mathbf{y})\, d\gamma(\mathbf{y}) \end{aligned}$$

are defined for $\mathbf{x} \in \Gamma$. We recall that

$$V_k^+ \in \mathcal{L}\Big(H^{-1/2}(\Gamma), H^{1/2}(\Gamma)\Big) \quad , \quad K_k^+ \in \mathcal{L}\Big(H^{-1/2}(\Gamma), H^{1/2}(\Gamma)\Big),$$
$$\big(K_k^+\big)^t \in \mathcal{L}\Big(H^{-1/2}(\Gamma), H^{1/2}(\Gamma)\Big) \quad , \quad D_k^+ \in \mathcal{L}\Big(H^{1/2}(\Gamma), H^{-1/2}(\Gamma)\Big).$$

Note that (16) is only one possible representation of u^+. The reason for using (16) is that the associated integral equation, namely (17), is also the integral equation for the resolution of the interior impedance problem. We set

$$\tilde{\mathbb{C}} := \begin{cases} \mathbb{C}\backslash\{z \in \mathbb{R}, z \leq 0\} & \text{if } N \text{ is even,} \\ \mathbb{C} & \text{otherwise.} \end{cases}$$

LEMMA 4. *The operator $R^+ \in \mathcal{L}\big(H^{-1/2}(\Gamma), H^1(\Omega)\big)$, defined for $k \in \mathcal{C}^+$ by Lemma 2, has a meromorphic extension to $\tilde{\mathbb{C}}$, and the extension is an operator belonging to the space $\mathcal{L}\big(H^{-1/2}(\Gamma), H^1_{loc}(\Omega)\big)$.*

Proof. The inverse of the operator T^+ is labeled S^+. We have that T^+ belongs to the space $\mathcal{L}\big(H^{1/2}(\Gamma), H^{-1/2}(\Gamma)\big)$ and $S^+ \in \mathcal{L}\big(H^{-1/2}(\Gamma), H^{1/2}(\Gamma)\big)$ in $\mathcal{C}^+$. The three operators R^+, T^+, and S^+ are well-defined in $\mathcal{C}^+$. Moreover, by (16) and (17), R^+ and S^+ are linked by the relation

$$R^+(k, \zeta(k)) = \Big(-i\tilde{V_k^+}\zeta(k) - \tilde{K_k^+}\Big) S^+(k, \zeta(k)). \tag{18}$$

Since $\Big(-i\tilde{V_k^+}\zeta(k) - \tilde{K_k^+}\Big)$ is analytic in $\tilde{\mathbb{C}}$, it is equivalent to extend R^+ or S^+. Henceforth the extension of R^+ is done through T^+. First, by (17), one can see that T^+ is well-defined $\forall\ k \in \tilde{\mathbb{C}}$ and is analytic in k. Then, from [5], we notice that for k_0, such that $\Im(k_0) > 0$, the operator $D_{k_0}^+$ is invertible. We write for $k \in \tilde{\mathbb{C}}$

$$T^+(k, \zeta(k)) = -D_{k_0}^+ [\mathrm{I} + C(k, \zeta(k))],$$

where

$$\begin{aligned} C(k, \zeta(k)) := {D_{k_0}^+}^{-1}\big(D_k^+ - D_{k_0}^+\big) + i{D_{k_0}^+}^{-1}\zeta(k)K_k^+ \\ + i{D_{k_0}^+}^{-1}\big(K_k^+\big)^t\zeta(k) - {D_{k_0}^+}^{-1}\zeta(k)V_k^+\zeta(k). \end{aligned}$$

By subtracting the symbol of the pseudodifferential operators D_k^+ and $D_{k_0}^+$, one can easily show that ${D_{k_0}^+}^{-1}\big(D_k^+ - D_{k_0}^+\big)$ is compact (see [3, 14]). Moreover, from the mapping properties of ${D_{k_0}^+}^{-1}$, K_k^+, $\big(K_k^+\big)^t$, and V_k^+, the three remaining pseudodifferential operators appearing in the expression of $C(k, \zeta(k))$ are compact. We conclude that $C(k, \zeta(k))$ is compact. Finally by Lemma 2, $T^+(k, \zeta(k))$ is invertible $\forall\ k \in \mathcal{C}^+$. Consequently, by Steinberg's theorem [17], $\big(T^+(k, \zeta(k))\big)^{-1} = S^+(k, \zeta(k))$ has a meromorphic extension to $\tilde{\mathbb{C}}\backslash\mathcal{C}^+$. Hence, S^+ has a countable number of poles. Thanks to relation (18) and to the analycity of $\zeta(k)$ and of the two operators $\tilde{V_k^+}$ and $\tilde{K_k^+}$, R^+ is extended in the same way as S^+. ∎

Remark 5. The proof of Lemma 4 gives the way to define the outgoing solution of the Helmholtz equation for almost any frequency $k \in \tilde{\mathbb{C}}$. From now on, we will refer to this procedure when stating that a solution is outgoing.

As a consequence of the above two lemmas, we have the following theorem.

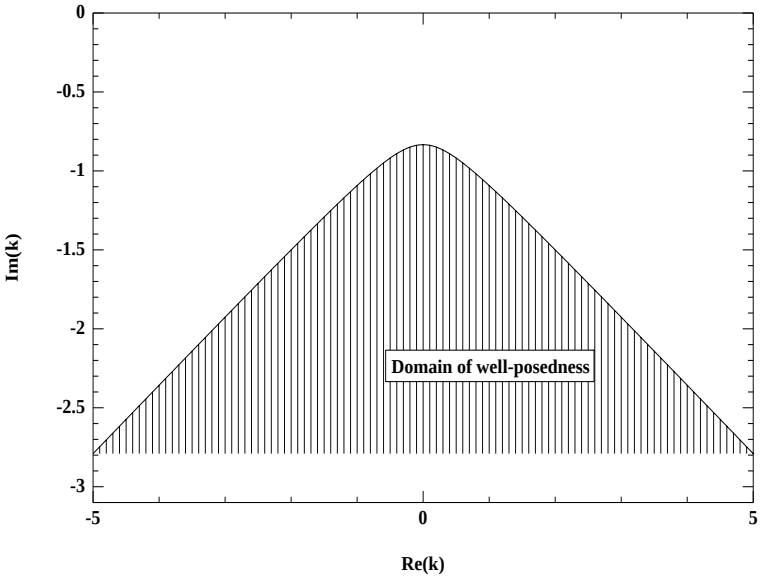


FIG. 2. *The set* $\mathcal{C}^-$.

THEOREM 6. *R^+ has a countable number of poles. These poles lie in $\tilde{\mathbb{C}}\backslash\mathcal{C}^+$.* Conversely, the incoming solution u^- is defined for $\Im(k) < 0$ by

$$\left\{\begin{array}{l} u^- \in H^1(\Omega)\,, \\ -\Delta u^- - k^2 u^- = 0 \quad \text{in} \quad \Omega, \\ \frac{\partial u^-}{\partial \mathbf{n}} + i\zeta(k) u^- = g \quad \text{on} \quad \Gamma. \end{array}\right. \tag{19}$$

As for the outgoing problem, one can show the following lemma.

LEMMA 7. *Let $\zeta(k)$ be an impedance function satisfying condition* (12)*. Then the problem* (19) *has a unique solution when the complex frequency k lies inside the set $\mathcal{C}^-$ defined by*

$$\Im(k) \leq -\lambda_m |k| - (\lambda_m C + 1) \quad \text{and} \quad |k| \geq K,$$

where $C > 0$ depends only on the domain Ω (see Figure 2*).*

Remark 8. When $\Im(k) < 0$, the requirement that the solution to the Helmholtz equation belongs to $H^1(\Omega)$ is equivalent to the incoming Sommerfeld condition [4]. Namely, whenever $\Im(k) < 0$, (19) is equivalent to the following problem:

$$\left\{\begin{array}{l} u^- \in H^1_{loc}(\Omega)\,, \\ -\Delta u^- - k^2 u^- = 0 \quad \text{in} \quad \Omega, \\ \frac{\partial u^-}{\partial \mathbf{n}} + i\zeta(k) u^- = g \quad \text{on} \quad \Gamma, \\ \lim_{r\to\infty} r^{\frac{N-1}{2}} \left(\frac{\partial u^-}{\partial r} + iku^-\right) = 0 \quad \text{uniformly in all directions.} \end{array}\right.$$

This defines the *"incoming"* solution of the Helmholtz equation.

We define the operator $R^-(k, \zeta(k))$ by $R^-(k, \zeta(k))\, g = u^-$ for $k \in \mathcal{C}^-$. All the previous integral operators can be defined for the incoming problem. The superscript $+$ has to be replaced by $-$. The incoming fundamental solution $G_k^-(\mathbf{x}, \mathbf{y})$ is equal to

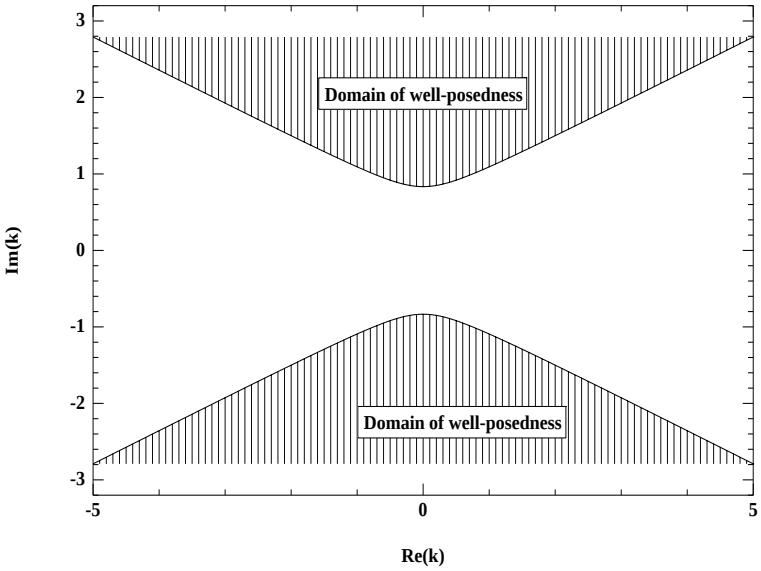


FIG. 3. *The set* $\mathcal{C}^{\text{int}}$.

$-\frac{i}{4}H_0^{(2)}(k|\mathbf{x}-\mathbf{y}|)$ for $N=2$ and $\frac{\exp(-ik|\mathbf{x}-\mathbf{y}|)}{4\pi|\mathbf{x}-\mathbf{y}|}$ for $N=3$. Equation (17) with the superscript $-$ instead of $+$ is the integral equation for the incoming problem and defines the operator T^-. The operators R^- and S^- are extended in a meromorphic way to $\mathbb{C}\backslash\mathcal{C}^-$, leading to the following lemma.

LEMMA 9. *The operator* $R^- \in \mathcal{L}\big(H^{-1/2}(\Gamma), H^1(\Omega)\big)$ *defined for* $k \in \mathcal{C}^-$ *by Lemma* 7 *has a meromorphic extension to* $\tilde{\mathbb{C}}$, *and the extension is an operator belonging to* $\mathcal{L}\big(H^{-1/2}(\Gamma), H^1_{loc}(\Omega)\big)$.

THEOREM 10. R^- *has a countable number of poles. These poles lie in* $\tilde{\mathbb{C}}\backslash\mathcal{C}^-$.

Finally, it appears useful to define the impedance interior problem,

$$\left\{\begin{array}{ll} u^{\text{int}} \in H^1\big(\Omega^{\text{int}}\big), & \\ -\Delta u^{\text{int}} - k^2 u^{\text{int}} = 0 & \text{in} \quad \Omega, \\ \frac{\partial u^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k) u^{\text{int}} = g & \text{on} \quad \Gamma, \end{array}\right. \tag{20}$$

and the operator $R^{\text{int}}(k,\zeta(k))$ by $R^{\text{int}}(k,\zeta(k))\, g = u^{\text{int}}$ when u^{int} is well-defined. Similarly to Lemmas 2 and 7, we have the following.

LEMMA 11. *Let* $\zeta(k)$ *be an impedance function satisfying condition* (12). *The problem* (20) *has a unique solution when the complex frequency* k *lies inside the set* $\mathcal{C}^{\text{int}}$ *defined by* (*see Figure* 3)

$$|\Im(k)| \geq \lambda_m |k| + \lambda_m C + 1 \text{ and } |k| \geq K,$$

where $C > 0$ *depends only on the domain* Ω.

The solution u^{int} can be represented by an integral representation involving either the outgoing kernel G_k^+ or the incoming kernel G_k^-:

$$R^{\text{int}} = \left(-i\tilde{V_k^+}\zeta(k) - \tilde{K_k^+}\right) S^+(k,\zeta(k)),$$

or

$$R^{\rm int} = \left(-i\tilde{V_k^-}\zeta(k) - \tilde{K_k^-}\right) S^-(k,\zeta(k))\,.$$

Obviously, these two expressions are equivalent inside $\Omega^{\rm int}$. As in Lemmas 4 and 9, one can show that the problem (20) has a countable number of poles.

THEOREM 12. *$R^{\rm int}$ has a countable number of poles. These poles lie in $\tilde{\mathbb{C}}\backslash\mathcal{C}^{\rm int}$.*

The importance of the interior problem for the construction of the scattering matrix comes from the following two relations,

$$S^+(k,\zeta(k)) : g \mapsto v^+ = [\![u^+]\!] := {u^{\rm int}}_{|\Gamma} - {u^+}_{|\Gamma},$$
$$S^-(k,\zeta(k)) : g \mapsto v^- = [\![u^-]\!] := {u^{\rm int}}_{|\Gamma} - {u^-}_{|\Gamma},$$

where u^+, u^- and $u^{\rm int}$ are the solutions of (13), (19), and (20), respectively, with the right-hand side g.

2.2. Link between incoming and outgoing solutions. In fact, the two problems (13) and (19) are very close to one another. Thus, special relations link R^+ and R^-. This is what makes the relation (9) possible. First, if $u^+ = R^+(k,\zeta(k))\,g$ for $k \in \mathcal{C}^+$, we see that the incoming Sommerfeld condition holds with $(-k)$ instead of k. Then, since $-k \in \mathcal{C}^-$, by Lemma 7, we get

$$R^+(k,\zeta(k)) = R^-(-k,\zeta(k)) \tag{21}$$

whenever $k \in \mathcal{C}^+$. By analyticity arguments, this relation holds $\forall\ k \in \tilde{\mathbb{C}}$ such that $R^+(k,\zeta(k))$ is defined. This shows that the poles of $R^+(k,\zeta(k))$ are exactly those of $R^-(-k,\zeta(k))$.

On the other hand, if one sets $u = R^+(k,\zeta(k))\,\overline{g}$, then $\overline{u}$ satisfies for $k \in \mathcal{C}^+$

$$\begin{cases} -\Delta\overline{u} - \overline{k}^2\overline{u} = 0 & \text{in} \quad \Omega, \\ \frac{\partial\overline{u}}{\partial\mathbf{n}} - i\,\overline{\zeta(k)}\overline{u} = g & \text{on} \quad \Gamma, \\ \lim_{r\to\infty} r^{\frac{N-1}{2}}\left(\frac{\partial\overline{u}}{\partial r} + i\overline{k}\overline{u}\right) = 0. \end{cases}$$

By Lemma 7, this problem has a unique solution since $\overline{k} \in \mathcal{C}^-$ and the couple $\left(\overline{k}, -\overline{\zeta(k)}\right)$ satisfies (12). Hence $\overline{u} = R^-\left(\overline{k}, -\overline{\zeta(k)}\right) g$. Then

$$\overline{R^+(k,\zeta(k))} = R^-\left(\overline{k}, -\overline{\zeta(k)}\right) \tag{22}$$

for $k \in \mathcal{C}^+$. Using the same argument as for (21), relation (22) holds for almost all $k \in \tilde{\mathbb{C}}$. With the same argument, we also have

$$R^{\rm int}(k,\zeta(k)) = R^{\rm int}(-k,\zeta(k))\,, \qquad \overline{R^{\rm int}(k,\zeta(k))} = R^{\rm int}\left(\overline{k}, -\overline{\zeta(k)}\right). \tag{23}$$

If $T^\pm$ stands for either T^+ or T^-, we have

$$\begin{aligned} \left\langle T^\pm(k,\zeta(k))\,v, v'\right\rangle = &-\int_{\Gamma\times\Gamma} \frac{\partial^2}{\partial\mathbf{n_x}\partial\mathbf{n_y}} G_k^\pm(\mathbf{x},\mathbf{y})\,v(\mathbf{x})\overline{v}'(\mathbf{y})\,d\gamma(\mathbf{x})\;d\gamma(\mathbf{y}) \\ -i\int_{\Gamma\times\Gamma} &\left[\frac{\partial}{\partial\mathbf{n_y}} G_k^\pm(\mathbf{x},\mathbf{y})\,\zeta(k,\mathbf{x}) + \frac{\partial}{\partial\mathbf{n_x}} G_k^\pm(\mathbf{x},\mathbf{y})\,\zeta(k,\mathbf{y})\right] v(\mathbf{x})\overline{v}'(\mathbf{y})\,d\gamma(\mathbf{x})\;d\gamma(\mathbf{y}) \\ +\int_{\Gamma\times\Gamma} &G_k^\pm(\mathbf{x},\mathbf{y})\,\zeta(k,\mathbf{x})\,\zeta(k,\mathbf{y})\,v(\mathbf{x})\overline{v}'(\mathbf{y})\,d\gamma(\mathbf{x})\;d\gamma(\mathbf{y})\,. \end{aligned}$$

One can easily check the relation $\overline{G_k^{\pm}(\mathbf{x},\mathbf{y})} = G_{-\overline{k}}^{\pm}(\mathbf{x},\mathbf{y})$ for both the incoming and the outgoing solutions. Let $(T^{\pm}(k,\zeta(k)))^*$ be the adjoint operator of $T^{\pm}(k,\zeta(k))$. Since the kernel of $T^{\pm}$ is symmetric with respect to $\mathbf{x}$ and $\mathbf{y}$, we have $(T^{\pm}(k,\zeta(k)))^* = \overline{T^{\pm}(k,\zeta(k))}$, and thus

$$\left\langle \left(T^{\pm}(k,\zeta(k))\right)^* v, v' \right\rangle = \left\langle \overline{T^{\pm}(k,\zeta(k))} v, v' \right\rangle = \left\langle T^{\pm}\left(-\overline{k}, -\overline{\zeta(k)}\right) v, v' \right\rangle$$

for almost all $k \in \tilde{\mathbb{C}}$. Hence

$$\left(T^{\pm}(k,\zeta(k))\right)^* = \overline{T^{\pm}(k,\zeta(k))} = T^{\pm}\left(-\overline{k}, -\overline{\zeta(k)}\right) \tag{24}$$

for almost all $k \in \tilde{\mathbb{C}}$. If $S^{\pm}$ stands for S^+ or S^-, then we also have

$$\left(S^{\pm}(k,\zeta(k))\right)^* = \overline{S^{\pm}(k,\zeta(k))} = S^{\pm}\left(-\overline{k}, -\overline{\zeta(k)}\right) \tag{25}$$

for almost all $k \in \tilde{\mathbb{C}}$. Similarly, the relation $G_{-k}^{+}(\mathbf{x},\mathbf{y}) = G_k^{-}(\mathbf{x},\mathbf{y})$ enables us to show the following equality

$$S^-(k,\zeta(k)) = S^+(-k,\zeta(k)) \tag{26}$$

for almost all $k \in \tilde{\mathbb{C}}$.

2.3. The far field patterns. The outgoing and incoming solutions share a special behavior at infinity (i.e., very far away from the obstacle): the solution at a point $\mathbf{x} \in \mathbb{R}^N$ is asymptotically equal to a function of $|\mathbf{x}|$ times a function of $\hat{\mathbf{x}} := \frac{\mathbf{x}}{|\mathbf{x}|}$. This latter function is called the far field pattern. The main interest of the far field for what we are concerned with lies in the fact that it gives a characterization of the scattering matrix.

The outgoing kernel satisfies

$$G_k^+(\mathbf{x},\mathbf{y}) = \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N-1}{2}}} e^{-ik\hat{\mathbf{x}}\cdot\mathbf{y}} \chi(k) + O\left(\frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N+1}{2}}}\right), \qquad \text{where } \chi(k) = \frac{(-ik)^{\frac{N-3}{2}}}{2(2\pi)^{\frac{N-1}{2}}},$$

and

$$\frac{\partial G_k^+}{\partial \mathbf{n_y}}(\mathbf{x},\mathbf{y}) = \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N-1}{2}}} \frac{\partial e^{-ik\hat{\mathbf{x}}\cdot\mathbf{y}}}{\partial \mathbf{n_y}} \chi(k) + O\left(\frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N+1}{2}}}\right).$$

Recalling that $v^+ = S^+(k,\zeta(k))\, g = [\![u^+]\!]$ and using (18), we get

$$R^+(k,\zeta(k))\, g(\mathbf{x}) = \frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N-1}{2}}} A_{+,g}(\hat{\mathbf{x}},k,\zeta(k)) + O\left(\frac{e^{ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N+1}{2}}}\right), \tag{27}$$

where for any $\hat{\mathbf{x}} \in S_1$

$$A_{+,g}(\hat{\mathbf{x}},k,\zeta(k)) := -\chi(k) \int_\Gamma \left(\frac{\partial e^{-ik\hat{\mathbf{x}}\cdot\mathbf{y}}}{\partial \mathbf{n_y}} + i\zeta(k,\mathbf{y}) e^{-ik\hat{\mathbf{x}}\cdot\mathbf{y}}\right) [\![u^+(\mathbf{y})]\!]\, d\gamma(\mathbf{y}).$$

For the incoming solution, we have

$$G_k^-(\mathbf{x},\mathbf{y}) = \frac{e^{-ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N-1}{2}}} e^{ik\hat{\mathbf{x}}\cdot\mathbf{y}} \chi(-k) + O\left(\frac{e^{-ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N+1}{2}}}\right)$$

and

$$\frac{\partial G_k^-}{\partial \mathbf{n_y}}(\mathbf{x},\mathbf{y}) = \frac{e^{-ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N-1}{2}}}\frac{\partial e^{ik\hat{\mathbf{x}}\cdot\mathbf{y}}}{\partial \mathbf{n_y}}\chi(-k) + O\left(\frac{e^{-ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N+1}{2}}}\right).$$

Hence

$$R^-(k,\zeta(k))\, g(\mathbf{x}) = \frac{e^{-ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N-1}{2}}}A_{-,g}(\hat{\mathbf{x}},k,\zeta(k)) + O\left(\frac{e^{-ik|\mathbf{x}|}}{|\mathbf{x}|^{\frac{N+1}{2}}}\right), \tag{28}$$

where for any $\hat{\mathbf{x}} \in S_1$

$$A_{-,g}(\hat{\mathbf{x}},k,\zeta(k)) := -\chi(-k)\int_\Gamma \left(\frac{\partial e^{ik\hat{\mathbf{x}}\cdot\mathbf{y}}}{\partial \mathbf{n_y}} + i\zeta(k,\mathbf{y})e^{ik\hat{\mathbf{x}}\cdot\mathbf{y}}\right)[\![u^-(\mathbf{y})]\!]\, d\gamma(\mathbf{y}).$$

$A_{+,g}$ and $A_{-,g}$ are the scattering amplitudes, or far field patterns of, respectively, the outgoing and the incoming solution.

From (21), we automatically infer that

$$A_{-,g}(\hat{\mathbf{x}},k,\zeta(k)) = A_{+,g}(\hat{\mathbf{x}},-k,\zeta(k)), \tag{29}$$

and (22) implies that

$$\overline{A_{+,g}(\hat{\mathbf{x}},k,\zeta(k))} = A_{-,\overline{g}}\left(\hat{\mathbf{x}},\overline{k},-\overline{\zeta(k)}\right). \tag{30}$$

Now we study more closely the case of the plane wave incidence. This is important for several reasons. First, it is a very important case physically. Second, the far field pattern can be expressed as a superposition of plane waves. More precisely, for the incident wave $u_0(\mathbf{x}) = e^{-ik\hat{\mathbf{y}}\cdot\mathbf{x}}$ where $|\hat{\mathbf{y}}| = 1$, the boundary condition reads

$$\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u = \frac{\partial u_0}{\partial \mathbf{n}} + i\zeta(k)u_0 = q_{\hat{\mathbf{y}},k,\zeta(k)},$$

where

$$\left(q_{\hat{\mathbf{y}},k,\zeta(k)}\right)(\mathbf{x}) := \frac{\partial}{\partial \mathbf{n_x}}e^{-ik\hat{\mathbf{y}}\cdot\mathbf{x}} + i\zeta(k,\mathbf{x})e^{-ik\hat{\mathbf{y}}\cdot\mathbf{x}}. \tag{31}$$

The far field pattern, with respect to the plane wave, is labeled $s_\pm$, i.e.,

$$s_\pm(\hat{\mathbf{y}},\hat{\mathbf{x}},k,\zeta(k)) := A_{\pm,q_{\hat{\mathbf{y}},k,\zeta(k)}}(\hat{\mathbf{x}},k,\zeta(k)). \tag{32}$$

Putting together (27), (28), and (31), we have that

$$\begin{aligned} A_{+,g}(\hat{\mathbf{x}},k,\zeta(k)) &= -\chi(k)\int_\Gamma q_{\hat{\mathbf{x}},k,\zeta(k)}\, S^+(k,\zeta(k))\, g\, d\gamma \\ A_{-,g}(\hat{\mathbf{x}},k,\zeta(k)) &= -\chi(-k)\int_\Gamma q_{-\hat{\mathbf{x}},k,\zeta(k)}\, S^-(k,\zeta(k))\, g\, d\gamma. \end{aligned} \tag{33}$$

2.4. Two fundamental relationships between incoming and outgoing far field patterns.

LEMMA 13. *We have*

$$\begin{aligned} A_{-,g}(-\hat{\mathbf{x}},k,\zeta(k)) &= -1^{\frac{N+1}{2}} A_{+,g}(\hat{\mathbf{x}},k,\zeta(k)) \\ &\quad + \frac{ik}{2\pi}^{\frac{N-1}{2}} \int_{|\hat{\mathbf{y}}|=1} s_-(\hat{\mathbf{x}},\hat{\mathbf{y}},k,\zeta(k))\, A_{+,g}(\hat{\mathbf{y}},k,\zeta(k))\, dS(\hat{\mathbf{y}}), \end{aligned} \tag{34}$$

$$\begin{aligned} A_{+,g}(\hat{\mathbf{x}},k,\zeta(k)) &= -1^{\frac{N+1}{2}} A_{-,g}(-\hat{\mathbf{x}},k,\zeta(k)) \\ &\quad + \left(\frac{-ik}{2\pi}\right)^{\frac{N-1}{2}} \int_{|\hat{\mathbf{y}}|=1} s_+(\hat{\mathbf{x}},-\hat{\mathbf{y}},k,\zeta(k))\, A_{-,g}(-\hat{\mathbf{y}},k,\zeta(k))\, dS(\hat{\mathbf{y}}). \end{aligned} \tag{35}$$

Proof. For $k \in \tilde{\mathbb{C}}$ and $g \in H^{-1/2}(\Gamma)$ fixed, let us compute

$$\begin{aligned} \mathcal{H} &= -\frac{A_{+,g}(\hat{\mathbf{x}},k,\zeta(k))}{\chi(k)} + \frac{A_{-,g}(-\hat{\mathbf{x}},k,\zeta(k))}{\chi(-k)} \\ &= \int_\Gamma \left[S^+(k,\zeta(k))\, g - S^-(k,\zeta(k))\, g\right] q_{\hat{\mathbf{x}},k,\zeta(k)}\, d\gamma. \end{aligned}$$

Thanks to (26), one may write

$$\mathcal{H} = \left\langle S^+(k,\zeta(k))\, g, \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} \right\rangle - \left\langle g, \left(S^+(-k,\zeta(k))\right)^* \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} \right\rangle.$$

Due to (25), we arrange the previous equation as

$$\mathcal{H} = \left\langle S^+(k,\zeta(k))\, g, \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} \right\rangle - \left\langle g, S^+\left(\overline{k}, -\overline{\zeta(k)}\right) \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} \right\rangle.$$

By (31), the plane wave satisfies $q := \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} = q_{-\hat{\mathbf{x}},\overline{k},-\overline{\zeta(k)}}$. We define the function v by $v = R^+\left(\overline{k}, -\overline{\zeta(k)}\right) q$ in Ω and by $v = R^{\text{int}}\left((\overline{k}, -\overline{\zeta(k)}\right) q$ in Ω^{int}. In fact, v is the outgoing solution of $\Delta v + \overline{k}^2 v = 0$. Moreover, we have $S^+\left(\overline{k}, -\overline{\zeta(k)}\right)\overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} = S^+\left(\overline{k}, -\overline{\zeta(k)}\right) q = [\![v]\!]$ and $\frac{\partial v}{\partial \mathbf{n}} - i\overline{\zeta(k)} v = q = \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}}$ on both sides of Γ. From (27) and (32), the asymptotic behavior of v is

$$v(\mathbf{y}) \overset{|\mathbf{y}|\to\infty}{\sim} \frac{e^{i\overline{k}|\mathbf{y}|}}{|\mathbf{y}|^{\frac{N-1}{2}}} s_+\left(-\hat{\mathbf{x}}, \hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\right),$$

where s_+ is defined in (32).

On the other hand, let us denote by $u = R^+(k,\zeta(k))\, g$ in Ω and $u = R^{\text{int}}(k,\zeta(k))\, g$ in Ω^{int}, the outgoing solution with the initial data g. We have

$$S^+(k,\zeta(k))\, g = [\![u]\!], \qquad \frac{\partial u}{\partial \mathbf{n}} + i\zeta(k) u = g \quad \text{on} \quad \Gamma$$

and

$$u(\mathbf{y}) \overset{|\mathbf{y}|\to\infty}{\sim} \frac{e^{ik|\mathbf{y}|}}{|\mathbf{y}|^{\frac{N-1}{2}}} A_{+,g}(\hat{\mathbf{y}},k,\zeta(k)).$$

Therefore

$$\begin{aligned}\mathcal{H} &= \left\langle [\![u]\!], \frac{\partial v}{\partial \mathbf{n}} - i\overline{\zeta(k)}v \right\rangle - \left\langle \frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u, [\![v]\!] \right\rangle \\ &= \int_\Gamma \left\{ [\![u]\!] \left(\frac{\partial \overline{v}}{\partial \mathbf{n}} + i\zeta(k)\overline{v} \right) - \left(\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u \right) [\![\overline{v}]\!] \right\}.\end{aligned}$$

Let us denote by $\tilde{\Omega}_R$ the domain $\Omega^{\text{int}} \cup (\Omega \cap B_R)$ for some $R > 0$ large enough. Then Green's formula applied to $\tilde{\Omega}_R$ turns into the two following relations:

$$\begin{aligned}\int_{\tilde{\Omega}_R} \nabla u \cdot \nabla \overline{v} - k^2 u\overline{v} &= \int_{\tilde{\Omega}_R} \nabla u \cdot \nabla \overline{v} + \Delta u\, \overline{v} \\ &= \int_\Gamma \left(\frac{\partial u^{\text{int}}}{\partial \mathbf{n}} \overline{v}^{\text{int}} - \frac{\partial u}{\partial \mathbf{n}} \overline{v} \right) d\gamma + \int_{S_R} \frac{\partial u}{\partial r} \overline{v}\, dS_R\end{aligned}$$

and

$$\begin{aligned}\int_{\tilde{\Omega}_R} \nabla u \cdot \nabla \overline{v} - k^2 u\overline{v} &= \int_{\tilde{\Omega}_R} \nabla u \cdot \nabla \overline{v} + u\Delta \overline{v} \\ &= \int_\Gamma \left(u^{\text{int}} \frac{\partial \overline{v}^{\text{int}}}{\partial \mathbf{n}} - u \frac{\partial \overline{v}}{\partial \mathbf{n}} \right) d\gamma + \int_{S_R} u \frac{\partial \overline{v}}{\partial r}\, dS_R.\end{aligned}$$

By subtracting these last two equations, we have

$$\begin{aligned}&- \int_{S_R} \left(\frac{\partial u}{\partial r} \overline{v} - u \frac{\partial \overline{v}}{\partial r} \right) dS_R \\ &= \int_\Gamma \left\{ \left(\frac{\partial u^{\text{int}}}{\partial \mathbf{n}} \overline{v}^{\text{int}} - \frac{\partial u}{\partial \mathbf{n}} \overline{v} \right) - \left(u^{\text{int}} \frac{\partial \overline{v}^{\text{int}}}{\partial \mathbf{n}} - u \frac{\partial \overline{v}}{\partial \mathbf{n}} \right) \right\} d\gamma \\ &= \int_\Gamma \left\{ \left[\left(\frac{\partial u^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k)u^{\text{int}} \right) \overline{v}^{\text{int}} - \left(\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u \right) \overline{v} \right] \right. \\ &\quad - \left. \left[u^{\text{int}} \left(\frac{\partial \overline{v}^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k)\overline{v}^{\text{int}} \right) - u \left(\frac{\partial \overline{v}}{\partial \mathbf{n}} + i\zeta(k)\overline{v} \right) \right] \right\} d\gamma.\end{aligned}$$

Since

$$\frac{\partial u^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k)u^{\text{int}} = \frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u$$

and

$$\frac{\partial \overline{v}^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k)\overline{v}^{\text{int}} = \frac{\partial \overline{v}}{\partial \mathbf{n}} + i\zeta(k)\overline{v},$$

the last equality becomes

$$\int_{S_R} \left(\frac{\partial u}{\partial r} \overline{v} - u \frac{\partial \overline{v}}{\partial r} \right) dS_R = \int_\Gamma \left\{ [\![u]\!] \left(\frac{\partial \overline{v}}{\partial \mathbf{n}} + i\zeta(k)\overline{v} \right) - \left(\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u \right) [\![\overline{v}]\!] \right\} d\gamma. \tag{36}$$

Now on S_R, the asymptotic behavior of u and v enables one to write

$$\frac{\partial u(r\hat{\mathbf{y}})}{\partial r} \overline{v(r\hat{\mathbf{y}})} - u(r\hat{\mathbf{y}}) \frac{\partial \overline{v(r\hat{\mathbf{y}})}}{\partial r} \overset{r\to\infty}{\sim} \frac{2ik}{r^{N-1}} \overline{s_+\left(-\hat{\mathbf{x}}, \hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\right)} A_{+,g}(\hat{\mathbf{y}}, k, \zeta(k)).$$

Thus,

$$\mathcal{H} = \int_{S_R} \left(\frac{\partial u}{\partial r} \overline{v} - u \frac{\partial \overline{v}}{\partial r} \right) \, dS_R$$

with

$$\int_{S_R} \left(\frac{\partial u}{\partial r} \overline{v} - u \frac{\partial \overline{v}}{\partial r} \right) \, dS_R \overset{R\to\infty}{\sim} \frac{2ik}{R^{N-1}} \int_{S_R} \overline{s_+\Big(-\hat{\mathbf{x}}, \hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\Big)} A_{+,g}(\hat{\mathbf{y}}, k, \zeta(k)) \, dS_R.$$

By (30) and (32)

$$\begin{aligned} \overline{s_+\Big(-\hat{\mathbf{x}}, \hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\Big)} &= \overline{A_{+,q_{-\hat{\mathbf{x}},\overline{k},-\overline{\zeta(k)}}}\Big(\hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\Big)} \\ &= \overline{A_{+,\overline{q_{\hat{\mathbf{x}},k,\zeta(k)}}}\Big(\hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\Big)} \\ &= A_{-,q_{\hat{\mathbf{x}},k,\zeta(k)}}(\hat{\mathbf{y}}, k, \zeta(k)) = s_-(\hat{\mathbf{x}}, \hat{\mathbf{y}}, k, \zeta(k)) \, . \end{aligned}$$

Thus,

$$\overline{s_+\Big(-\hat{\mathbf{x}}, \hat{\mathbf{y}}, \overline{k}, -\overline{\zeta(k)}\Big)} = s_-(\hat{\mathbf{x}}, \hat{\mathbf{y}}, k, \zeta(k)) \tag{37}$$

and

$$\int_{S_R} \left(\frac{\partial u}{\partial r} \overline{v} - u \frac{\partial \overline{v}}{\partial r} \right) \, dS_R \overset{R\to\infty}{\longrightarrow} 2ik \int_{|\hat{\mathbf{y}}|=1} s_-(\hat{\mathbf{x}}, \hat{\mathbf{y}}, k, \zeta(k)) \, A_{+,g}(\hat{\mathbf{y}}, k, \zeta(k)) \, dS(\hat{\mathbf{y}}).$$

Since $\mathcal{H}$ does not depend on R, it follows that

$$\mathcal{H} = 2ik \int_{|\hat{\mathbf{y}}|=1} s_-(\hat{\mathbf{x}}, \hat{\mathbf{y}}, k, \zeta(k)) \, A_{+,g}(\hat{\mathbf{y}}, k, \zeta(k)) \, dS(\hat{\mathbf{y}}).$$

Finally from the relations $\frac{\chi(k)}{\chi(-k)} = (-1)^{\frac{N-3}{2}} = (-1)^{\frac{N+1}{2}}$ and $2ik\chi(-k) = \left(\frac{ik}{2\pi}\right)^{\frac{N-1}{2}}$, we have proved (34).

Let us now explicitly write down the mapping $A_- \to A_+$. We proceed as previously. Thanks to (25) and (26), $\mathcal{H}$ is given by

$$\begin{aligned} \mathcal{H} &= -\int_\Gamma \left\{ S^-(k, \zeta(k)) \, g - S^-(-k, \zeta(k)) \, g \right\} q_{\hat{\mathbf{x}},k,\zeta(k)} \, d\gamma \\ &= -\left\langle S^-(k, \zeta(k)) \, g, \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} \right\rangle + \left\langle g, S^-\Big(\overline{k}, -\overline{\zeta(k)}\Big) \, \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} \right\rangle . \end{aligned}$$

We now set

$$\begin{cases} v = R^-\Big(\overline{k}, -\overline{\zeta(k)}\Big) \, \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} & \text{in } \Omega \\ v = R^{\text{int}}\Big(\overline{k}, -\overline{\zeta(k)}\Big) \, \overline{q_{\hat{\mathbf{x}},k,\zeta(k)}} & \text{in } \Omega^{\text{int}} \end{cases} \quad \text{and} \quad \begin{cases} u = R^-(k, \zeta(k)) \, g & \text{in } \Omega \\ u = R^{\text{int}}(k, \zeta(k)) \, g & \text{in } \Omega^{\text{int}}. \end{cases}$$

Hence

$$\mathcal{H} = \int_\Gamma \left\{ \left(\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k) u \right) [\![\overline{v}]\!] - [\![u]\!] \left(\frac{\partial \overline{v}}{\partial \mathbf{n}} + i\zeta(k) \overline{v} \right) \right\} .$$

Formula (36) also holds with the new choice of u and v. Thus,

$$\begin{aligned}\mathcal{H} &= -\int_{S_R}\left(\frac{\partial u}{\partial r}\overline{v} - u\frac{\partial \overline{v}}{\partial r}\right) dS_R \\ &= 2ik\int_{|\hat{\mathbf{y}}|=1} \overline{s_-\Big(-\hat{\mathbf{x}},\hat{\mathbf{y}},\overline{k},-\overline{\zeta(k)}\Big)}A_{-,g}(\hat{\mathbf{y}},k,\zeta(k))\,dS(\hat{\mathbf{y}}).\end{aligned}$$

Finally we have proved (35) by (37) and by changing $\hat{\mathbf{y}}$ into $-\hat{\mathbf{y}}$ in the integral. ∎

3. The scattering matrix. In section 2, we have introduced several notions in the case of a general impedance boundary condition. We are going to apply them to the case of an impedance that satisfies the physical assumption

$$0 \le \Re\big(\zeta(k,\mathbf{x})\overline{k}\big) \le \lambda_m\,|k|^2 \quad \forall k\in\mathbb{C} \text{ with } |k|\ge K \quad \text{and} \quad \forall \mathbf{x}\in\Gamma \tag{38}$$

for some $K \ge 0$ and $0 < \lambda_m < 1$. We use the notation of section 2.

3.1. Notion of resonant frequency. The calculations of section 2.1 are also valid in the case of (38). However few results can be improved. In particular, Lemma 2 becomes the following.

LEMMA 14. *Let $\zeta(k)$ be an impedance function satisfying condition* (38)*. The problem* (13) *has a unique solution when the complex frequency k lies inside the upper half plane $\mathbb{C}^+ := \{k\in\mathbb{C}\ ,\ \Im(k) > 0\}$.*

Proof. We proceed as in the proof of Lemma 2. From (14) and (38), one can automatically deduce that $i\overline{k}b^+$ is coercive when $\Im(k) > 0$. We conclude the proof by using the Lax–Milgram lemma. ∎

As in section 2.1, we define $R^+(k,\zeta(k)) : g \mapsto u^+$, where u^+ is solution to (13). Using Lemma 4, one can extend R^+ to $\tilde{\mathbb{C}}\backslash\mathbb{C}^+$. Hence R^+ has a countable number of poles in $\tilde{\mathbb{C}}\backslash\mathbb{C}^+$.

DEFINITION 15. *The "resonant frequencies" are defined as the poles of the operator $R^+(k,\zeta(k))$, where $\zeta(k)$ is a physical impedance.*

Combining Lemmas 4 and 14, we get the following theorem.

THEOREM 16. *The resonant frequencies lie in $\tilde{\mathbb{C}}\backslash\mathbb{C}^+$.*

3.2. Definition of the scattering matrix and derivation of the fundamental relation. In [9], the scattering matrix for the Dirichlet or the Neumann boundary condition is defined from the problem in time. However, in [10, formula (A.16)], it is shown that the scattering matrix satisfies a relation which involves only the far fields of the outgoing and incoming solutions. For the impedance boundary condition, if the scattering matrix is denoted by $\mathcal{S}(k,\zeta(k))$, this relation writes formally

$$\mathcal{S}(k,\zeta(k)) \ :\ A_{-,g}(-\hat{\mathbf{x}},k,\zeta(k)) \ \mapsto\ A_{+,g}(\hat{\mathbf{x}},k,\zeta(k)) \tag{39}$$

$\forall\ g \in H^{-1/2}(\Gamma)$. In order to infer from this property a definition of the scattering matrix, we must get rid of the term g in (39). To this end, let us introduce the far field operators $A_\pm(k,\zeta(k))$ defined by

$$\big(A_\pm(k,\zeta(k))\,g\big)(\hat{\mathbf{x}}) = A_{\pm,g}(\hat{\mathbf{x}},k,\zeta(k))\,.$$

If W stands for the symmetry operator (i.e., $[Wa]\,(\hat{\mathbf{y}}) = a(-\hat{\mathbf{y}})$), then (39) turns into

$$\mathcal{S}(k,\zeta(k))\,W A_-(k,\zeta(k))\,g = A_+(k,\zeta(k))\,g$$

$\forall\ g \in H^{-1/2}(\Gamma)$. Hence $\mathcal{S}(k,\zeta(k))\, W A_-(k,\zeta(k)) = A_+(k,\zeta(k))$, which implies the following definition.

DEFINITION 17. *The "scattering matrix" $\mathcal{S}(k,\zeta(k))$ associated with the frequency k and the impedance $\zeta(k)$ is defined by*

$$\mathcal{S}(k,\zeta(k)) := A_+(k,\zeta(k))\,(A_-(k,\zeta(k)))^{-1}\, W^{-1}.$$

The operators $A_\pm(k,\zeta(k))$ map $H^{-1/2}(\Gamma)$ to $L^2(S_1)$. We will show in the proof of Theorem 18 that $A_\pm(k,\zeta(k))$ is one-to-one. Hence $A_\pm(k,\zeta(k))$ is an isomorphism from $H^{-1/2}(\Gamma)$ onto the space $A_\pm(k,\zeta(k))H^{-1/2}(\Gamma)$. Actually, the space $A_\pm(k,\zeta(k))H^{-1/2}(\Gamma)$ is strictly included in $L^2(S_1)$. We do not know how to characterize the far field patterns of waves scattered by an obstacle in $L^2(S_1)$ [5]. Hence the scattering matrix is only defined on $WA_\pm(k,\zeta(k))H^{-1/2}(\Gamma)$. However, we will show later on that the scattering matrix can be extended to any function of $L^2(S_1)$.

THEOREM 18. *The poles of the scattering matrix are the poles of the far field operator $A_+(k,\zeta(k))$.*

Proof. From Definition 17, the poles of the scattering matrix are composed of the poles of $A_+(k,\zeta(k))$ plus the zeros of $A_-(k,\zeta(k))$. Thus the property is proved if we show that there is no zero of $A_-(k,\zeta(k))$.

Suppose by contradiction that k is a zero of $A_-(k,\zeta(k))$. Then 0 is eigenvalue of $A_-(k,\zeta(k))$. Hence there exists a nonzero function $g \in H^{-1/2}(\Gamma)$ such that the far field $A_-(k,\zeta(k))\, g := u^\infty$ of the incoming solution u to the extension of the problem (19) vanishes (see Lemma 9). Hence from [5, Theorem 2.13], we have $u \equiv 0$ outside a ball containing the obstacle Ω^{int}, and thus by analytic continuation (see [4]) $u \equiv 0$ in Ω. Taking the restriction of u on the boundary, we get $\frac{\partial u}{\partial \mathbf{n}} + i\zeta(k)u = 0$ on Γ. Finally, since $g \not\equiv 0$, the impedance boundary condition cannot be satisfied, which leads to a contradiction. ∎

3.3. Expression of the scattering matrix. We set

$$[Q_+(k,\zeta(k))\, A]\,(\hat{\mathbf{x}}) = \int_{|\hat{\mathbf{y}}|=1} s_+(\hat{\mathbf{x}},-\hat{\mathbf{y}},k,\zeta(k))\, A(\hat{\mathbf{y}}) dS(\hat{\mathbf{y}}),$$
$$[Q_-(k,\zeta(k))\, A]\,(\hat{\mathbf{x}}) = \int_{|\hat{\mathbf{y}}|=1} s_-(\hat{\mathbf{x}},\hat{\mathbf{y}},k,\zeta(k))\, A(\hat{\mathbf{y}}) dS(\hat{\mathbf{y}}).$$

By (35) and (39), the scattering matrix writes

$$\mathcal{S}(k,\zeta(k)) = (-1)^{\frac{N+1}{2}} \left[\mathrm{I} - \left(\frac{ik}{2\pi}\right)^{\frac{N-1}{2}} Q_+(k,\zeta(k)) \right]. \tag{40}$$

From this expression of the scattering matrix, it is now clear that the domain of definition of $\mathcal{S}$ can be extended to the whole space $L^2(S_1)$. Thanks to (34), the inverse of $\mathcal{S}$ is given by

$$\mathcal{S}^{-1}(k,\zeta(k)) = (-1)^{\frac{N+1}{2}} \left[\mathrm{I} - \left(\frac{-ik}{2\pi}\right)^{\frac{N-1}{2}} Q_-(k,\zeta(k)) \right]. \tag{41}$$

The conjugate of the scattering matrix is

$$\mathcal{S}^*(k,\zeta(k)) = (-1)^{\frac{N+1}{2}} \left[\mathrm{I} - \left(\frac{-i\bar{k}}{2\pi}\right)^{\frac{N-1}{2}} Q_+(k,\zeta(k))^* \right].$$

The kernel of $Q_+(k,\zeta(k))$ is $s_+(\hat{\mathbf{x}},-\hat{\mathbf{y}},k,\zeta(k))$, so that the kernel of $Q_+(k,\zeta(k))^*$ is $\overline{s_+(\hat{\mathbf{y}},-\hat{\mathbf{x}},k,\zeta(k))}$. From the reciprocity relation [5, 9], we have $s_+(\hat{\mathbf{y}},-\hat{\mathbf{x}},k,\zeta(k)) = s_+(-\hat{\mathbf{x}},\hat{\mathbf{y}},k,\zeta(k))$. Thus, by relation (37), $s_-\big(\hat{\mathbf{x}},\hat{\mathbf{y}},\overline{k},-\overline{\zeta(k)}\big)$ is the kernel of the operator $Q_+(k,\zeta(k))^*$, which implies that $Q_+((k,\zeta(k))^* = Q_-\big(\overline{k},-\overline{\zeta(k)}\big)$. Hence

$$\mathcal{S}^*(k,\zeta(k)) = \mathcal{S}^{-1}\Big(\overline{k},-\overline{\zeta(k)}\Big)\,,$$

and we obtain the fundamental relation

$$\mathcal{S}\Big(\overline{k},-\overline{\zeta(k)}\Big) = [\mathcal{S}^*(k,\zeta(k))]^{-1}\,. \tag{42}$$

Hence relation (9) is shown. We have insisted in the introduction on the fact that this relation plays a central role in this work. We now give a simple consequence of (42).

It is important for applications to see whether or not the scattering matrix is unitary on the real axis. Indeed, the scattering matrix is unitary on the real axis if and only if the time translation operator $U(t)$ is unitary. This latter property means that the energy of the systems is conserved. Because of the surface impedance which introduces some absorption, we expect the scattering matrix to be not unitary. The answer to this question is given in the next lemma.

LEMMA 19. *The scattering matrix is unitary on the real axis if and only if* $\zeta(k)$ *is purely imaginary on the real axis.*

As a consequence of this lemma, the assumption (11) implies that the scattering matrix is not unitary on the real axis. In fact, if (11) is not satisfied, then the impedance boundary condition degenerates into the Robin boundary condition on the real axis.

Proof. The scattering matrix is unitary on the real axis if and only if

$$\mathcal{S}^*(k,\zeta(k)) = [\mathcal{S}(k,\zeta(k))]^{-1}$$

for almost all $k \in \mathbb{R}\cap\tilde{\mathbb{C}}$. Hence by (42), the scattering matrix is unitary on the real axis if and only if $\mathcal{S}(k,\zeta(k)) = \mathcal{S}\big(k,-\overline{\zeta(k)}\big)$. By (40), this is equivalent to $s_+(\hat{\mathbf{x}},-\hat{\mathbf{y}},k,\zeta(k)) = s_+\big(\hat{\mathbf{x}},-\hat{\mathbf{y}},k,-\overline{\zeta(k)}\big)$ for almost all $\hat{\mathbf{x}} \in S_1$ and almost all $\hat{\mathbf{y}} \in S_1$. From the uniqueness of the recovery of the impedance (on a known and fixed boundary) from the far field pattern (see Theorem 6.13 in [4]), we conclude that the scattering matrix is unitary on the real axis if and only if $\zeta(k) = -\overline{\zeta(k)}$ for k real. ∎

3.4. Relation between the resonant frequencies and the zeros of the scattering matrix. The main interest of the scattering matrix lies in the following result.

THEOREM 20. *The poles of the scattering matrix* $\mathcal{S}(k,\zeta(k))$ *are exactly the resonant frequencies* (*see Definition* 15).

Proof. From Theorem 18 and relation (33), the poles of the scattering matrix are the poles of the jump operator $S^+(k,\zeta(k))$. Since $S^+(k,\zeta(k))\,g$ corresponds to the jump between the outgoing solution $R^+(k,\zeta(k))\,g$ and the interior solution $R^{\text{int}}(k,\zeta(k))\,g$, the poles of the scattering matrix are a priori composed of the poles of $R^+(k,\zeta(k))$ (namely the resonant frequencies) and the poles of R^{int}. We are going to show that the poles of the interior problem disappear so that the theorem is proved.

From (40), to show this we have only to show that the kernel s_+ of the operator Q_+ is independent of the interior problem. We recall that the interior problem appears in the jump operator S^+, i.e.,

$$s_+(\hat{\mathbf{x}},\hat{\mathbf{y}},k,\zeta(k)) = A_{+,q_{\hat{\mathbf{x}},k,\zeta(k)}}(\hat{\mathbf{y}},k,\zeta(k))$$

$$= -\chi(k) \int_\Gamma q_{\hat{\mathbf{y}},k,\zeta(k)} S^+(k,\zeta(k))\, q_{\hat{\mathbf{x}},k,\zeta(k)}\, d\gamma$$
$$= -\chi(k) \int_\Gamma q_{\hat{\mathbf{y}},k,\zeta(k)} \left(u^{\text{int}} - u^+\right)\, d\gamma,$$

where u^{int} is solution to (20) with $g = q_{\hat{\mathbf{x}},k,\zeta(k)}$. We set $e_{\hat{\mathbf{x}}}(\mathbf{y}) = e^{-ik\,\hat{\mathbf{x}}.\mathbf{y}}$. We notice that $\forall\ k \in \mathbb{C}$, $e_{\hat{\mathbf{x}}}$ is solution to (20) with $g = q_{\hat{\mathbf{x}},k,\zeta(k)}$. When k is not a pole of (20), we have uniqueness of (20) and thus $u^{\text{int}}(\mathbf{y}) = e_{\hat{\mathbf{x}}}(\mathbf{y})$. Now when k is a pole of (20), we do not want to single out the particular solution $e_{\hat{\mathbf{x}}}$. We assume that u^{int} is any solution to (20). Green's formula then gives

$$0 = \int_{\Omega^{\text{int}}} e_{\hat{\mathbf{y}}} \Delta u^{\text{int}} - \Delta e_{\hat{\mathbf{y}}}\, u^{\text{int}} = \int_\Gamma e_{\hat{\mathbf{y}}} \frac{\partial u^{\text{int}}}{\partial \mathbf{n}} - \frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} u^{\text{int}}\, d\gamma.$$

Thus, since $\frac{\partial u^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k) u^{\text{int}} = \frac{\partial e_{\hat{\mathbf{x}}}}{\partial \mathbf{n}} + i\zeta(k) e_{\hat{\mathbf{x}}}$ on Γ, it follows that

$$\int_\Gamma q_{\hat{\mathbf{y}},k,\zeta(k)} u^{\text{int}}\, d\gamma = \int_\Gamma e_{\hat{\mathbf{y}}} \left(\frac{\partial u^{\text{int}}}{\partial \mathbf{n}} + i\zeta(k) u^{\text{int}}\right)\, d\gamma = \int_\Gamma e_{\hat{\mathbf{y}}} \left(\frac{\partial e_{\hat{\mathbf{x}}}}{\partial \mathbf{n}} + i\zeta(k) e_{\hat{\mathbf{x}}}\right)\, d\gamma.$$

We can also show that, by Green's formula,

$$\int_\Gamma e_{\hat{\mathbf{y}}} \frac{\partial e_{\hat{\mathbf{x}}}}{\partial \mathbf{n}}\, d\gamma = \int_\Gamma \frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} e_{\hat{\mathbf{x}}}\, d\gamma.$$

Thus,

$$\int_\Gamma q_{\hat{\mathbf{y}},k,\zeta(k)} u^{\text{int}}\, d\gamma = \int_\Gamma \left(\frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} + i\zeta(k) e_{\hat{\mathbf{y}}}\right) e_{\hat{\mathbf{x}}}\, d\gamma = \int_\Gamma q_{\hat{\mathbf{y}},k,\zeta(k)} e_{\hat{\mathbf{x}}}\, d\gamma$$

and

$$s_+(\hat{\mathbf{x}},\hat{\mathbf{y}},k,\zeta(k)) = -\chi(k) \int_\Gamma q_{\hat{\mathbf{y}},k,\zeta(k)} \left(e_{\hat{\mathbf{x}}} - u^+\right)\, d\gamma. \tag{43}$$

We notice that the interior problem disappears: the spurious poles (of the interior problem) are not poles of $\mathcal{S}$. □

Remark 21. The scattering matrix gives the relationship between incoming and outgoing solutions. In order to give an explicit form of $\mathcal{S}$ in the previous section, we used integral equations on the boundary of the obstacle. The main drawback to the use of integral equations is that they introduce the interior problem and its singularities in the formulation of $\mathcal{S}$. In particular, we know that the problem (20) is singular for a countable number of poles. Theorem 20 shows that these poles do not create spurious poles for $\mathcal{S}$, that is to say, the poles of the scattering matrix are in fact the resonant frequencies of the exterior problem.

THEOREM 22. *k is a resonant frequency (i.e., a pole of $\mathcal{S}$ for the impedance $\zeta(k)$) if and only if $\overline{k}$ is a zero of $\mathcal{S}$ for the impedance $-\overline{\zeta(k)}$.*

Proof. From (42), k is a pole of $\mathcal{S}^*$ for the impedance $\zeta(k)$ if and only if $\overline{k}$ is a zero of $\mathcal{S}$ for the impedance $-\overline{\zeta(k)}$. To conclude, we use Theorem 20 and Lemma 23 below. □

LEMMA 23. *Let $B(k)$ be an operator-valued meromorphic function. Then $B(k)$ and $B(k)^*$ have their poles at the same complex numbers k.*

Proof. Assume that k be a pole of $B(k)$. Let k_n be a sequence which converge to k. Then there exists a sequence u_n such that $\|u_n\| = 1$ and $\|B(k_n)u_n\|$ tends to infinity. We set $v_n = B(k_n)u_n$. We have

$$\|v_n\| = \left(B(k_n)u_n \, , \, \frac{v_n}{\|v_n\|} \right) = \left(u_n \, , \, B(k_n)^* \frac{v_n}{\|v_n\|} \right) \leq \|u_n\| \left\| B(k_n)^* \frac{v_n}{\|v_n\|} \right\|.$$

We conclude that $\left\| B(k_n)^* \frac{v_n}{\|v_n\|} \right\|$ tends to infinity, which implies that k is a pole of $B(k)^*$. ∎

4. Purely imaginary resonant frequencies. In this section, some estimates on the location of the purely imaginary poles are given. Thanks to Theorem 22, we are able to use the work of P. Lax and R. Phillips [9, 10] and J. Beale [2]. Here we are concerned with an impedance of the form

$$\zeta(k, \mathbf{x}) = k\lambda(\mathbf{x}), \tag{44}$$

where $0 \leq \lambda(\mathbf{x}) < 1$ for any $\mathbf{x} \in \Gamma$. Relation (11) is satisfied. Moreover, $\lambda_m := \max_{\mathbf{x}\in\Gamma} \lambda(\mathbf{x}) < 1$, which implies that (12) holds. Condition (38) is trivially satisfied so that all the results of previous section hold with the choice (44). Equation (44) is the impedance function which is most broadly used in the literature.

We first give the expression of the scattering matrix on the purely imaginary axis. It is of the form (4) where $Q(i\sigma)$ is a compact self-adjoint operator when $\sigma \in \mathbb{R}$.

4.1. The transmission coefficient. As in [10, 2], we wish to study the location of the zeros of $\mathcal{S}$ instead of the location of the poles of $\mathcal{S}$.

THEOREM 24. *The resonant frequencies are the poles of the extension of the problem* (13) *with the impedance* (44)*. These resonant frequencies lie in the lower half plane* $\tilde{\mathbb{C}}\backslash\mathbb{C}^+$*.*

Moreover, $-i\sigma$ *(with* $\sigma \in \mathbb{R}$*) is a resonant frequency if and only if* 0 *is an eigenvalue of the operator* $\mathcal{S}(i\sigma, -i\lambda\sigma)$*.*

Proof. The first part of the theorem is a simple consequence of Definition 15 and Theorem 16.

By Theorem 22, $k = -i\sigma$ is a resonant frequency (i.e., a pole of $\mathcal{S}$ for the impedance $\zeta(k) = -i\lambda\sigma$) if and only if $\overline{k} = i\sigma$ is a zero of $\mathcal{S}$ for the impedance $-\overline{\zeta(k)} = -i\lambda\sigma$. ∎

As stated in the introduction, we do not consider, for the study of the zeros of $\mathcal{S}$, the problem (13) at the frequency $k = i\sigma$ with the impedance (44), i.e., $\zeta(i\sigma) = i\sigma\lambda$. We consider instead the problem (13) at the frequency $k = i\sigma$ with the impedance $-i\sigma\lambda$. Thus we have to find the location of the zeros of the scattering matrix $\mathcal{S}\big(i\sigma, \tilde{\zeta}(i\sigma)\big)$, with the impedance $\tilde{\zeta}(k, x) = -\lambda(x)k$. From (40),

$$\mathcal{S}(i\sigma, -i\sigma\lambda) = (-1)^{\frac{N+1}{2}} \left[\mathrm{I} - \left(\frac{-\sigma}{2\pi} \right)^{\frac{N-1}{2}} s(\sigma) W \right], \tag{45}$$

where the operator W is defined by the relation $[Wa](\hat{\mathbf{y}}) = a(-\hat{\mathbf{y}})$, and where $s(\sigma)$ is the operator of $L^2(S_1)$ whose kernel is $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma) := s_+(\hat{\mathbf{x}}, \hat{\mathbf{y}}, i\sigma, -i\lambda\sigma)$. By (27) and (32), $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma)$ is given by

$$s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma) = \lim_{r\to\infty} \left[r^{\frac{N-1}{2}} e^{\sigma\rho} v(\rho\hat{\mathbf{y}}, \hat{\mathbf{x}}; \sigma) \right],$$

where

$$\left\{ \begin{array}{l} \left(\sigma^2 - \Delta\right) v(\mathbf{y}, \hat{\mathbf{x}}; \sigma) = 0 \quad \text{in} \quad \Omega, \\ \left(\frac{\partial}{\partial \mathbf{n}} + \lambda(\mathbf{y})\sigma\right) v(\mathbf{y}, \hat{\mathbf{x}}; \sigma) = \left(\frac{\partial}{\partial \mathbf{n}} + \lambda(\mathbf{y})\sigma\right) e^{\sigma\, \hat{\mathbf{x}}\cdot\mathbf{y}} \quad \text{on} \quad \Gamma. \end{array} \right. \tag{46}$$

This problem is elliptic and has a real-valued solution, so that $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma)$ is real valued. Moreover, from the reciprocity relation [5, 9], we have $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma) = s(\hat{\mathbf{y}}, \hat{\mathbf{x}}; \sigma)$. Consequently, the operator $s(\sigma)$ is self-adjoint. That $s(\sigma)$ is compact follows from the fact that $s(\sigma)$ is a Hilbert–Schmidt operator. $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma)$ is called the *"transmission coefficient"*.

Let us now finish this section by giving an integral representation of the transmission coefficient s. To this end, we set $v_{\hat{\mathbf{x}}}(\mathbf{z}) = v(\mathbf{z}, \hat{\mathbf{x}}; \sigma)$ and $e_{\hat{\mathbf{y}}}(\mathbf{z}) = e^{\sigma \hat{\mathbf{y}}\cdot\mathbf{z}}$.

LEMMA 25. *For $\sigma > 0$, the transmission coefficient is given by*

$$s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma) = \frac{\sigma^{\frac{N-3}{2}}}{2\,(2\pi)^{\frac{N-1}{2}}} \left[-\int_{\Omega^{\text{int}}} p_\sigma(e_{\hat{\mathbf{x}}}, e_{\hat{\mathbf{y}}}) - \int_\Omega p_\sigma(v_{\hat{\mathbf{x}}}, v_{\hat{\mathbf{y}}}) - \int_\Gamma \lambda\sigma(e_{\hat{\mathbf{x}}} e_{\hat{\mathbf{y}}} - v_{\hat{\mathbf{x}}} v_{\hat{\mathbf{y}}})\, d\gamma \right],$$

where $p_\sigma(u, v) = \nabla u \cdot \nabla v + \sigma^2 uv$.

Proof. From (43), we have (with $k = i\sigma$, and $\zeta(k) = -i\lambda\sigma$)

$$s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma) = -\chi(i\sigma) \int_\Gamma \left(\frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} + \lambda\sigma e_{\hat{\mathbf{y}}} \right) \left(e_{\hat{\mathbf{x}}} - v_{\hat{\mathbf{x}}} \right)\, d\gamma,$$

where $\chi(i\sigma) = \frac{\sigma^{\frac{N-3}{2}}}{2(2\pi)^{\frac{N-1}{2}}}$. Since $\frac{\partial v_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} + \lambda\sigma v_{\hat{\mathbf{y}}} = \frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} + \lambda\sigma e_{\hat{\mathbf{y}}}$, we get

$$s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma) = -\frac{\sigma^{\frac{N-3}{2}}}{2(2\pi)^{\frac{N-1}{2}}} \left\{ \int_\Gamma \left(\frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} + \lambda\sigma e_{\hat{\mathbf{y}}} \right) e_{\hat{\mathbf{x}}}\, d\gamma - \int_\Gamma \left(\frac{\partial v_{\hat{\mathbf{y}}}}{\partial \mathbf{n}} + \lambda\sigma v_{\hat{\mathbf{y}}} \right) v_{\hat{\mathbf{x}}}\, d\gamma \right\}.$$

Let us set $L_\sigma u = \left(\sigma^2 - \Delta\right) u$. Since $L_\sigma v_{\hat{\mathbf{y}}} = 0$ and $L_\sigma e_{\hat{\mathbf{y}}} = 0$, Green's formula leads to

$$0 = \int_{\Omega^{\text{int}}} e_{\hat{\mathbf{x}}} L_\sigma e_{\hat{\mathbf{y}}} = \int_{\Omega^{\text{int}}} p_\sigma(e_{\hat{\mathbf{x}}}, e_{\hat{\mathbf{y}}}) - \int_\Gamma e_{\hat{\mathbf{x}}} \frac{\partial e_{\hat{\mathbf{y}}}}{\partial \mathbf{n}}\, d\gamma$$

and

$$0 = \int_\Omega v_{\hat{\mathbf{x}}} L_\sigma v_{\hat{\mathbf{y}}} = \int_\Omega p_\sigma(v_{\hat{\mathbf{x}}}, v_{\hat{\mathbf{y}}}) + \int_\Gamma v_{\hat{\mathbf{x}}} \frac{\partial v_{\hat{\mathbf{y}}}}{\partial \mathbf{n}}\, d\gamma.$$

Hence, putting these two expressions in $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma)$, we have proved the lemma. ∎

4.2. Properties of the zeros of $\mathcal{S}$. The problem (46) does not correspond to a physical impedance (i.e., the impedance does not satisfy the condition (38)). However, the impedance satisfies the condition (12).

THEOREM 26. *Assume that $0 < \lambda_m := \max_{\mathbf{x}\in\Gamma} \lambda(\mathbf{x}) < 1$. Then for*

$$\sigma \geq \sigma_0' := \frac{\lambda_m C + 1}{1 - \lambda_m},$$

(where C is a constant which depends only on Ω) the problem (46) has a unique solution in $H^1(\Omega)$.

Proof. We only have to use Lemma 2 with the frequency $k = i\sigma$ and the impedance $\tilde{\zeta}(k) = -\lambda k$. We check that this impedance satisfies the condition (12) with $K = 0$.

Hence there exists a constant C depending only on Ω such that the problem (46) has a unique solution whenever $\sigma \geq \lambda_m \sigma + \lambda_m C + 1$. ∎

In the remaining, an impedance function λ will be said to be admissible if $\lambda(\mathbf{x}) \geq 0$ $\forall\ \mathbf{x} \in \Gamma$ and if $0 < \lambda_m := \max_{\mathbf{x}\in\Gamma} \lambda(\mathbf{x}) < 1$.

LEMMA 27. *Assume that* $0 < \lambda_m := \max_{\mathbf{x}\in\Gamma} \lambda(\mathbf{x}) < 1$. *Then the operator* $s(\sigma)$ *is negative for*

$$\sigma \geq \sigma_0'' := \frac{\lambda_m C}{1 - {\lambda_m}^2},$$

where C *is the same constant as in Theorem* 26.

Proof. If we set

$$E_a(\mathbf{x}) = \int_{|\hat{\mathbf{y}}|=1} e^{\sigma\, \hat{\mathbf{y}}\cdot\mathbf{x}} a(\hat{\mathbf{y}}) dS(\hat{\mathbf{y}}) \quad \text{and} \quad V_a(\mathbf{x}) = \int_{|\hat{\mathbf{y}}|=1} v(\mathbf{x}, \hat{\mathbf{y}}; \sigma)\, a(\hat{\mathbf{y}}) dS(\hat{\mathbf{y}}),$$

then one can write

$$(s(\sigma)a, a) = \frac{\sigma^{\frac{N-3}{2}}}{2\,(2\pi)^{\frac{N-1}{2}}} \left[-\int_{\Omega^{\text{int}}} p_\sigma(E_a, E_a) - \int_\Omega p_\sigma(V_a, V_a) - \int_\Gamma \lambda\sigma \left({E_a}^2 - {V_a}^2 \right) d\gamma \right].$$

The only positive term in the right-hand side of above equation is $\int_\Gamma \lambda\sigma {V_a}^2\, d\gamma$. Using (15) with $\mathcal{D} = \Omega$ and $\epsilon = \frac{1}{\lambda_m \sigma}$, we have

$$\lambda_m \sigma \int_\Gamma {V_a}^2\, d\gamma \leq \int_\Omega |\nabla V_a|^2 + \lambda_m \sigma (C + \lambda_m \sigma) \int_\Omega {V_a}^2.$$

Therefore for $\sigma \geq \sigma_0''$, we have $\sigma^2 \geq \lambda_m \sigma (C + \lambda_m \sigma)$, which proves that

$$(s(\sigma)a, a) \leq \frac{\sigma^{\frac{N-3}{2}}}{2\,(2\pi)^{\frac{N-1}{2}}} \left[-\int_{\Omega^{\text{int}}} p_\sigma(E_a, E_a) - \int_\Gamma \lambda\sigma {E_a}^2\, d\gamma \right] \leq 0. \quad ∎$$

We set $\sigma_0 := \max(\sigma_0', \sigma_0'')$. For $\sigma \geq \sigma_0$, the above theorem shows that the eigenvalues of $s(\sigma)$ are all negative. But when $\sigma < \sigma_0$, some eigenvalues may be positive. More precisely, Beale showed in [2] that the number of positive eigenvalues counted with multiplicities is finite. In the same way, most of the results of [2] are also valid in our case of interest.

The kernel $s(\hat{\mathbf{x}}, \hat{\mathbf{y}}; \sigma)$ is square integrable. Thus $s(\sigma)$ is an Hilbert–Schmidt operator. $s(\sigma)$ is then compact, which implies that the eigenvalues of $s(\sigma)$ are bounded and that zero is the only possible cluster point.

As for Theorem 3.7 in [2], the next theorem shows that $s(\sigma)$ is an increasing function of the interior domain.

THEOREM 28. *Let us introduce two bounded domains* Ω_1^{int} *and* Ω_2^{int} *such that* $\Omega_1^{\text{int}} \subset \Omega_2^{\text{int}}$. *The associated admissible impedances are respectively denoted by* λ_1 *and* λ_2. *We set* $s_i(\sigma)$ *the transmission operator with respect to* Ω^{int}. *Then for* $\sigma \geq \sigma_0$

$$s_2(\sigma) \leq s_1(\sigma) \leq 0.$$

Proof. For $a \in L^2(S_1)$, we set

$$E(\mathbf{x}) = \int_{|\hat{\mathbf{y}}|=1} e^{\sigma\, \hat{\mathbf{y}}\cdot\mathbf{x}} a(\hat{\mathbf{y}}) dS(\hat{\mathbf{y}}), \qquad V_i(\mathbf{x}) = \int_{|\hat{\mathbf{y}}|=1} v_i(\mathbf{x}, \hat{\mathbf{y}}; \sigma)\, a(\hat{\mathbf{y}}) dS(\hat{\mathbf{y}}).$$

We set $\Omega_j = \mathbb{R}^N \backslash \overline{\Omega_i^{(j)}}$. Then

$$\begin{aligned}
&2\,(2\pi)^{\frac{N-1}{2}}\,\sigma^{\frac{3-N}{2}}\,[(s_1(\sigma)a,a)-(s_2(\sigma)a,a)]\\
&=\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}(p_\sigma(E,E)-p_\sigma(V_1,V_1))+\int_{\Omega_2}(p_\sigma(V_2,V_2)-p_\sigma(V_1,V_1))\\
&\quad-\int_{\Gamma_1}\lambda_1\sigma\left(E^2-{V_1}^2\right)\,d\gamma+\int_{\Gamma_2}\lambda_2\sigma\left(E^2-{V_2}^2\right)\,d\gamma\\
&=\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}p_\sigma(E-V_1,E-V_1)+\int_{\Omega_2}p_\sigma(V_2-V_1,V_2-V_1)\\
&\quad+2\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}p_\sigma(E-V_1,V_1)+2\int_{\Omega_2}p_\sigma(V_2-V_1,V_1)\\
&\quad-\int_{\Gamma_1}\lambda_1\sigma\left(E^2-{V_1}^2\right)\,d\gamma+\int_{\Gamma_2}\lambda_2\sigma\left(E^2-{V_2}^2\right)\,d\gamma.
\end{aligned}$$

Using Green's formula, we have that

$$\begin{aligned}
0=\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}L_\sigma(E-V_1)\,V_1&=\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}p_\sigma(E-V_1,V_1)\\
&\quad-\int_{\Gamma_1}\lambda_1\sigma\,(E-V_1)\,V_1\,d\gamma-\int_{\Gamma_2}\frac{\partial(E-V_1)}{\partial\mathbf{n}}V_1\,d\gamma
\end{aligned}$$

and

$$0=\int_{\Omega_2}L_\sigma(V_2-V_1)\,V_1=\int_{\Omega_2}p_\sigma(V_2-V_1,V_1)+\int_{\Gamma_2}\frac{\partial(V_2-V_1)}{\partial\mathbf{n}}V_1\,d\gamma.$$

Thus

$$\begin{aligned}
&\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}p_\sigma(E-V_1,V_1)+\int_{\Omega_2}p_\sigma(V_2-V_1,V_1)\\
&\qquad=\int_{\Gamma_1}\lambda_1\sigma\,(E-V_1)\,V_1\,d\gamma-\int_{\Gamma_2}\frac{\partial(V_2-E)}{\partial\mathbf{n}}V_1\,d\gamma\\
&\qquad=\int_{\Gamma_1}\lambda_1\sigma\,(E-V_1)\,V_1\,d\gamma+\int_{\Gamma_2}\lambda_2\sigma\,(V_2-E)\,V_1\,d\gamma.
\end{aligned}$$

Hence

$$\begin{aligned}
2\,(2\pi)^{\frac{N-1}{2}}\,\sigma^{\frac{3-N}{2}}\,[(s_1(\sigma)a,a)-(s_2(\sigma)a,a)]&=\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}p_\sigma(E-V_1,E-V_1)\\
&\quad+\int_{\Omega_2}p_\sigma(V_2-V_1,V_2-V_1)-\int_{\Gamma_1}\lambda_1\sigma\,(E-V_1)^2\,d\gamma\\
&\quad+\int_{\Gamma_2}\lambda_2\sigma\,(E-V_1)^2\,d\gamma-\int_{\Gamma_2}\lambda_2\sigma\,(V_1-V_2)^2\,d\gamma.
\end{aligned}$$

Only two terms are nonpositive in the right-hand side of this equation. By (15) with $\mathcal{D}=\Omega_1\cap\Omega_2^{\mathrm{int}}$ and $\epsilon=(\lambda_1)_m\,\sigma$, we have

$$-(\lambda_1)_m\,\sigma\int_{\Gamma_1}(E-V_1)^2\,d\gamma\geq(\lambda_1)_m\,\sigma\int_{\Gamma_2}(E-V_1)^2\,d\gamma-\int_{\Omega_1\cap\Omega_2^{\mathrm{int}}}|\nabla(E-V_1)|^2$$

$$-(\lambda_1)_m\,\sigma\,(C+(\lambda_1)_m\,\sigma)\int_{\Omega_1\cap\Omega_2^{\rm int}}|E-V_1|^2$$
$$\geq(\lambda_1)_m\,\sigma\int_{\Gamma_2}(E-V_1)^2\;d\gamma-\int_{\Omega_1\cap\Omega_2^{\rm int}}p_\sigma(E-V_1,E-V_1)$$

since $(\lambda_1)_m<1$ and $\sigma\geq\sigma_0$. Similarly, one can show that

$$-(\lambda_2)_m\,\sigma\int_{\Gamma_2}(V_1-V_2)^2\;d\gamma\geq-\int_{\Omega_2}p_\sigma(V_1-V_2,V_1-V_2)\,.$$

Therefore,

$$(s_1(\sigma)a,a)\geq(s_2(\sigma)a,a)\,.\qquad\square$$

As in [2, Corollary 3.8], we have the following corollary.

COROLLARY 29. *Suppose that in the above theorem* $\Omega_2^{\rm int}=\Omega^{\rm int}$ *is star-like with respect to the origin, and* $\Omega_1^{\rm int}=t\Omega^{\rm int}$, $0<t<1$. *We assume that* $\lambda_2(\mathbf{x})=\lambda(\mathbf{x})$ *and* $\lambda_1(\mathbf{x})=\frac{1}{t}\lambda\big(\frac{\mathbf{x}}{t}\big)$, $\mathbf{x}\in\Gamma$. *Then the number* σ_0 *in Theorem* 28 *can be chosen independent of* t.

Following [2, Theorem 3.9], we also have the following theorem.

THEOREM 30. *If* $\Omega^{\rm int}$ *is star-like, then* $\sigma^{\frac{N-1}{2}}s(\sigma)$ *is a decreasing function of* σ, *for* $\sigma\geq\sigma_0$.

We do not give a proof of this theorem since the proof done in [2] for the Robin boundary condition can clearly be extended to our case.

We turn now to study the eigenvalues of the scattering matrix (see (45)). Since the operator $s(\sigma)$ is compact, so is $K(\sigma):=-\big(\frac{-\sigma}{2\pi}\big)^{\frac{N-1}{2}}s(\sigma)W$. Therefore the eigenvalues have no cluster point, except possibly zero. In each compact set that does not include zero, there are at most a finite number of eigenvalues of $K(\sigma)$. Moreover, since $s(\hat{\mathbf{x}},\hat{\mathbf{y}};\sigma)$ is real, the eigenvalues of $s(\sigma)$ are real. Hence the eigenvalues of $K(\sigma)$ are real if N is odd. In this case, the positive eigenvalues μ_m and the negative ones κ_m of $K(\sigma)$ are labeled by

$$\mu_1\geq\mu_2\geq\cdots>0>\cdots\geq\kappa_2\geq\kappa_1,$$

where $\lim_{m\to\infty}\mu_m=\lim_{m\to\infty}\kappa_m=0$.

Our aim is to find the positive values of σ for which -1 is an eigenvalue of $K(\sigma)$. This problem turns into that of studying the negative eigenvalues $\kappa_m(\sigma)$ and picking out the values of σ such that $\kappa_m(\sigma)=-1$.

From the expression of $s(\hat{\mathbf{x}},\hat{\mathbf{y}};\sigma)$ stated in Lemma 25, it is easy to show that the operator $s(\sigma)$ is bounded and analytic when $\sigma\to0$, if $N>2$. When $N=2$, the factor $\sigma^{\frac{N-3}{2}}$ in the expression of s implies that $s(\sigma)$ tends to infinity when $\sigma\to0$. And for $N>2$, from the expression of $K(\sigma)$ we automatically can infer that $\lim_{\sigma\to0}K(\sigma)=0$, which proves that for any m, $\lim_{\sigma\to0}\mu_m(\sigma)=\lim_{\sigma\to0}\kappa_m(\sigma)=0$. With the same argument, we have $\kappa_m(\sigma)\overset{\sigma\to\infty}{\longrightarrow}-\infty$.

In fact, κ is an eigenvalue of $K(\sigma)$ if $\big(\pm\big(\frac{2\pi}{-\sigma}\big)^{\frac{N-1}{2}}\kappa\big)$ is an eigenvalue of $s(\sigma)$. The sign $\pm$ comes from the operator W. The presence of W in the expression of $K(\sigma)$ complicates the study of the eigenvalues and the eigenvectors of $K(\sigma)$. However, one can show, as in [2, Theorem 4.3], that the results on $s(\sigma)$ can be extended to $K(\sigma)$.

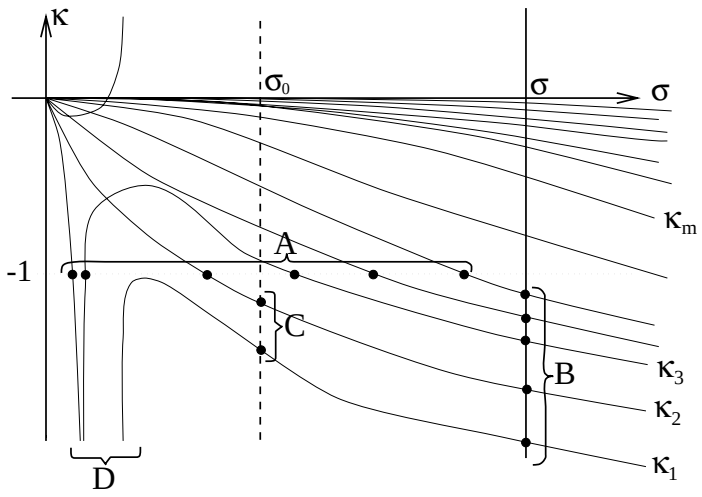


FIG. 4. *The function $\sigma \mapsto \kappa_m(\sigma)$ and its intersection with the line $\kappa = -1$. We have $A = N(0,\sigma;\Gamma,\lambda)$, $B = M(\sigma;\Gamma,\lambda)$, $C = M(\sigma_0;\Gamma,\lambda)$, and $D = N_{\mathrm{p}}(\Gamma,\Lambda)$.*

THEOREM 31. *Assume that $\sigma \geq \sigma_0$.*

- *If $\Omega_1^{\mathrm{int}} \subset \Omega_2^{\mathrm{int}}$ and if we denote by $\kappa_m^{(1)}$ and $\kappa_m^{(2)}$ (respectively, $\mu_m^{(1)}$ and $\mu_m^{(2)}$) the negative eigenvalues (respectively, the positive eigenvalues) with respect to Ω_1^{int} and Ω_2^{int}, then*

$$\kappa_m^{(2)}(\sigma) \leq \kappa_m^{(1)}(\sigma) \leq 0, \qquad \mu_m^{(2)}(\sigma) \geq \mu_m^{(1)}(\sigma) \geq 0.$$

- *$\kappa_m(\sigma)$ is a decreasing function of σ.*

We can also establish another result shown in [2, Theorem 4.4]: if N is even, then there are at most a finite number of zeros of $\mathcal{S}$. The reason for this comes from the relation

$$K(\sigma) = -i^{N-1}\Big(\frac{\sigma}{2\pi}\Big)^{\frac{N-1}{2}} s(\sigma)W.$$

We know by Lemma 27 that all the eigenvalues of $s(\sigma)$ are negative when $\sigma \geq \sigma_0$. In this case, the eigenvalues of $K(\sigma)$ are purely imaginary and thus cannot be equal to -1. On the other hand, if N is odd, there is an infinite sequence of zeros for each $\sigma > 0$. From now on, we assume that N is odd.

Let us denote by $N(\sigma_1,\sigma_2;\Gamma,\lambda)$ the number of zeros of the scattering matrix $\mathcal{S}(i\sigma,-i\lambda\sigma)$ when σ describes the slab $[\sigma_1,\sigma_2]$ (see Figure 4). Thanks to Theorem 31, it is quite clear what is going on for $\sigma \geq \sigma_0$: since $\kappa_m(\sigma)$ is decreasing and tends to minus infinity as $\sigma \to \infty$, then the equation $\kappa_m(\sigma) = -1$ has at most one solution $\sigma \geq \sigma_0$. Moreover, if $\kappa_m(\sigma_0) \geq -1$, then the equation $\kappa_m(\sigma) = -1$ has exactly one solution $\sigma \geq \sigma_0$. But in $[0,\sigma_0]$, $\kappa_m(\sigma)$ is not monotone and even can have some poles.

First let us show that the overall number of poles $\forall\ \kappa_m$ is bounded. $i\sigma$ is a pole of $\mathcal{S}(i\sigma,-i\lambda\sigma)$ if and only if σ is a pole of an eigenvalue κ_m. The set of all the poles of $\mathcal{S}(i\sigma,-i\lambda\sigma)$ has no cluster point, except at infinity. Furthermore, the multiplicity of each pole is finite. For any compact set of $\mathbb{C}$, there is at most a finite number of poles of $\mathcal{S}(i\sigma,-i\sigma\lambda)$. Therefore for $\sigma \in [0,\sigma_0]$ there is a finite number of poles and by Theorem 26, there are only a finite number of poles of $\mathcal{S}(i\sigma,-i\sigma\lambda)$ for $\sigma > 0$. The

finite number of poles for $\sigma > 0$ is denoted by $N_{\mathrm{p}}(\Gamma, \lambda)$. Each pole of the scattering matrix corresponds to a pole of one eigenvalue $\kappa_m(\sigma)$. We conclude that there exists an integer m_0 such that κ_m has no pole for $m \geq m_0$.

Using the same argument to the scattering matrix $\mathcal{S}(-i\sigma, -i\sigma\lambda)$ in the lower half plane, we conclude that this operator has only a finite number of poles in any compact set of $\mathbb{C}\backslash\{0\}$, and by Theorem 24 that $\mathcal{S}(i\sigma, -i\sigma\lambda)$ has a finite number of zeros in any finite compact set of $\sigma > 0$. To show that there is no cluster point at zero, we note that

$$\|K(\sigma)\| \leq \max(|\kappa_1(\sigma)|, \mu_1(\sigma)) \overset{\sigma\to 0}{\longrightarrow} 0.$$

Hence $\|K(\sigma)\| < 1$ in a compact set surrounding zero, which proves that $\mathcal{S}(i\sigma, -i\sigma\lambda)$ is invertible in this set. Therefore, zero cannot be a cluster point, and in the interval $[\sigma_1, \sigma_2]$ the number of zeros of $\mathcal{S}(i\sigma, -i\sigma\lambda)$ is finite for any $\sigma_1 \geq 0$. This number is clearly $N(\sigma_1, \sigma_2; \Gamma, \lambda)$.

LEMMA 32. *Assume that Ω^{int} is star-shaped. If $M(\sigma; \Gamma, \lambda)$ denotes the number of eigenvalues such that $\kappa_m(\sigma) \leq -1$ (see Figure 4), then for $\sigma > \sigma_0$*

$$N(\sigma_0, \sigma; \Gamma, \lambda) = M(\sigma; \Gamma, \lambda) - M(\sigma_0; \Gamma, \lambda)\,.$$

Moreover,

$$N(0, \sigma; \Gamma, \lambda) \geq M(\sigma; \Gamma, \lambda) - N_{\mathrm{p}}(\Gamma, \lambda)\,.$$

Proof. Since $\kappa_m(\sigma)$ is a decreasing function of σ for each m, we conclude that there are exactly $M(\sigma; \Gamma, \lambda) - M(\sigma_0; \Gamma, \lambda)$ eigenvalues κ_m such that $\kappa_m(\sigma_0) > -1$ and $\kappa_m(\sigma) \leq -1$. From Theorem 31, one can infer that each of these eigenvalues crosses the line -1 exactly once in $[\sigma_0, \sigma]$. The other eigenvalues cannot cross the straight line -1 in $[\sigma_0, \sigma]$, which proves the first equality.

For the second inequality, since $\kappa_m(0) = 0$ for any m, we notice that the $M(\sigma; \Gamma, \lambda)$ eigenvalues that are lower than -1 at the frequency σ must cross the line -1 at least once in $[0, \sigma]$ if and only if $\kappa_m(\sigma)$ is continuous in $[0, \sigma]$. The function $\kappa_m(\sigma)$ is continuous in $[0, \sigma]$ if there is no pole in $[0, \sigma]$. Since at most $N_{\mathrm{p}}(\Gamma, \lambda)$ eigenvalues have one pole or more, the theorem is proved. ∎

COROLLARY 33. *If $\Omega_1^{\mathrm{int}} \subset \Omega_2^{\mathrm{int}}$ with Ω_1^{int} star-shaped, and λ_1, λ_2 are two admissible impedances defined on Γ_1, Γ_2, respectively, then we have for $\sigma > \sigma_0$*

$$N(0, \sigma; \Gamma_2, \lambda_2) \geq N(\sigma_0, \sigma; \Gamma_1, \lambda_1) - N_{\mathrm{p}}(\Gamma_2, \lambda_2)\,.$$

Proof. From Lemma 32, we have

$$M := M(\sigma; \Gamma_1, \lambda_1) \geq N(\sigma_0, \sigma; \Gamma_1, \lambda_1)\,.$$

There are exactly M eigenvalues of Ω_1^{int} lower than -1 at the frequency σ, i.e.,

$$\kappa_1^{(1)}(\sigma) \leq -1, \ldots, \qquad \kappa_M^{(1)}(\sigma) \leq -1.$$

Thus by Theorem 31, $\kappa_1^{(2)}(\sigma) \leq -1$, …, $\kappa_M^{(2)}(\sigma) \leq -1$, which means that

$$M(\sigma; \Gamma_2, \lambda_2) \geq M.$$

Finally, by Lemma 32,

$$\begin{aligned} N(0, \sigma; \Gamma_2, \lambda_2) &\geq M(\sigma; \Gamma_2, \lambda_2) - N_{\mathrm{p}}(\Gamma_2, \lambda_2) \\ &\geq N(\sigma_0, \sigma; \Gamma_1, \lambda_1) - N_{\mathrm{p}}(\Gamma_2, \lambda_2)\,. \end{aligned}$$ ∎

THEOREM 34. *Assume that there exist two balls B_{R_1} and B_{R_2} of radii R_1 and R_2, respectively, such that $\overline{B}_{R_1} \subset \Omega^{\text{int}}$, $\overline{\Omega}_i \subset B_{R_2}$. Assume furthermore that Ω^{int} is star-shaped. Let λ, λ_1, and λ_2 be admissible impedances on Γ, S_{R_1}, and S_{R_2}, respectively. Then for $\sigma > \sigma_0$*

$$\begin{aligned} N(0,\sigma;\Gamma,\lambda) &\geq M(\sigma;S_{R_1},\lambda_1) - M(\sigma_0;S_{R_1},\lambda_1) - N_{\text{p}}(\Gamma,\lambda)\,, \\ N(0,\sigma;\Gamma,\lambda) &\leq M(\sigma;S_{R_2},\lambda_2) - M(\sigma_0;S_{R_2},\lambda_2) + N(0,\sigma_0;S_{R_2},\lambda_2) \\ &\quad + N_{\text{p}}(S_{R_2},\lambda_2) + N(0,\sigma_0;\Gamma,\lambda)\,. \end{aligned}$$

Proof. From Corollary 33, we have that

$$N(0,\sigma;\Gamma,\lambda) \geq N(\sigma_0,\sigma;S_{R_1},\lambda_1) - N_{\text{p}}(\Gamma,\lambda)\,.$$

Then by Lemma 32, the first inequality is proved.

Corollary 33 now enables us to write

$$N(0,\sigma;S_{R_2},\lambda_2) \geq N(\sigma_0,\sigma;\Gamma,\lambda) - N_{\text{p}}(S_{R_2},\lambda_2)\,.$$

Hence one may infer that

$$\begin{aligned} N(0,\sigma;\Gamma,\lambda) &= N(0,\sigma_0;\Gamma,\lambda) + N(\sigma_0,\sigma;\Gamma,\lambda) \\ &\leq N(0,\sigma;S_{R_2},\lambda_2) + N_{\text{p}}(S_{R_2},\lambda_2) + N(0,\sigma_0;\Gamma,\lambda)\,. \end{aligned}$$

By Lemma 32, we have

$$\begin{aligned} N(0,\sigma;S_{R_2},\lambda_2) &= N(0,\sigma_0;S_{R_2},\lambda_2) + N(\sigma_0,\sigma;S_{R_2},\lambda_2) \\ &= M(\sigma;S_{R_2},\lambda_2) - M(\sigma_0;S_{R_2},\lambda_2) + N(0,\sigma_0;S_{R_2},\lambda_2)\,. \end{aligned}$$

This, combined with previous relation, proves the second inequality of the theorem. ∎

In the right-hand side of the two inequalities of Theorem 34, only the first term depends on σ. If N is odd, these two terms tend to infinity as $\sigma \to \infty$, and hence gives the leading behavior of $N(0,\sigma;\Gamma,\lambda)$ for σ large. These two factors, namely $M(\sigma;S_{R_1},\lambda_1)$ and $M(\sigma;S_{R_2},\lambda_2)$ can be computed as in [2, section 5]. In fact, the asymptotic number of zeros on the imaginary axis for the case of the sphere is independent of the boundary condition. More precisely, the number of zeros on the imaginary axis between 0 and $i\sigma$ for the case of a sphere a radius R is asymptotically equal to $\frac{1}{(N-1)!}\left(\frac{R}{\gamma_0}\right)^{N-1}$, where $\gamma_0 \approx 0.6627$ [2]. Noting that $N_{\text{I}}(\sigma) = N(0,\sigma;\Gamma,\lambda)$, we can infer the following theorem.

THEOREM 35. *Assume N is odd. Assume further that $\overline{B}_{R_1} \subset \Omega^{\text{int}}$, $\overline{\Omega}_i \subset B_{R_2}$ and that Ω^{int} is star-shaped. Then*

$$\lim_{\sigma\to\infty} \frac{N_{\text{I}}(\sigma)}{\sigma^{N-1}} \geq \frac{1}{(N-1)!}\left(\frac{R_1}{\gamma_0}\right)^{N-1},$$

$$\lim_{\sigma\to\infty} \frac{N_{\text{I}}(\sigma)}{\sigma^{N-1}} \leq \frac{1}{(N-1)!}\left(\frac{R_2}{\gamma_0}\right)^{N-1},$$

where γ_0 is the unique solution of the equation

$$e^{\sqrt{1+\gamma_0{}^2}}\,\frac{\sqrt{1+\gamma_0{}^2}-1}{\gamma_0} = 1.$$

5. Conclusion. The motivation of this paper was to answer the following question: What can we infer of the obstacle from the location of the resonant frequencies? It is well known that the asymptotic repartition of eigenvalues of the interior acoustic problem gives the volume and the perimeter of the obstacle. This leads us to consider counting functions of the resonant frequencies for the exterior of an obstacle. It is not easy at all to link the global counting function with some geometrical quantities on the obstacle. However, a special study can be performed on the purely imaginary axis. There are infinitely many resonant frequencies on this axis. Let $N_{\mathrm{I}}(\sigma)$ be the number of purely imaginary poles whose modulus is lower than σ. We set

$$R_\Gamma = \gamma_0 \left[(N-1)! \lim_{\sigma \to \infty} \frac{N_{\mathrm{I}}(\sigma)}{\sigma^{N-1}} \right]^{\frac{1}{N-1}}.$$

Then for the acoustic waves with the impedance boundary condition (in odd space dimensions), we have

$$R_1 \leq R_\Gamma \leq R_2,$$

where R_1 is the radius of the largest sphere contained in the obstacle and R_2 is the radius of the smallest sphere containing the obstacle. Without further a priori information on the obstacle Ω^{int}, this relation is not of great interest for the inverse problem: it simply states that the boundary of the obstacle is crossing a sphere of radius R_Γ. This does not give a very precise idea of the size of the obstacle. For the inverse problem, it would be more interesting to have an upper bound on R_2 or a lower bound on R_1.

However, we are sometimes looking for an obstacle belonging to a special class of targets. For instance, for radar identification, one can look for an airplane. Since the shape of an airplane is always roughly the same, we can figure out that R_Γ gives a good idea of the size of the airplane. It is in this sense that the last inequality has to be used.

Acknowledgment. The author is grateful to David L. Colton for suggesting this subject and for many useful discussions.

REFERENCES

[1] C. Baum, *The Singularity Expansion Method*, in Transient Electromagnetic Fields, L.B. Felsen, ed., Springer-Verlag, New York, 1976, pp. 129–179.
[2] J. Beale, *Purely imaginary scattering frequencies for exterior domains*, Duke Math. J., 41 (1974), pp. 607–637.
[3] G. Chen, J. Zhou, *Boundary Element Methods*, Academic Press, London, 1992.
[4] D. Colton and R. Kress, *Integral Equation Methods in Scattering Methods in Scattering Theory*, Wiley-Interscience Publication, New York, 1983.
[5] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, Heidelberg, New York, 1992.
[6] I. Gohberg, M. Krein, *Introduction to the theory of the linear nonselfadjoint operators*, American Mathematical Society, Providence, RI, 1969.
[7] G. Grubb, *Functional Calculus of Pseudodifferential Boundary Problems*, Progress in Mathematics 65, Birkhauser, Boston, 1986.
[8] V. Isakov, *New stability results for soft obstacles in inverse scattering*, Inverse Problems, 9 (1993), pp. 535–543.
[9] P. Lax and R. Philipps, *Scattering Theory*, Academic Press, London, 1967.

[10] P. Lax and R. Philipps, *Decaying modes for the wave equation in the exterior of an obstacle*, Comm. Pure Appl. Math., XXII (1969), pp. 737–787.
[11] P. Lax and R. Phillips, *On the scattering frequencies of the Laplace operator for the exterior domain*, Comm. Pure Appl. Math., XXV (1972), pp. 85–101.
[12] P. Lax and R. Phillips, *Scattering theory for dissipative systems*, J. Funct. Anal., 14 (1973), pp. 172–235.
[13] P. Lax and R. Phillips, *A logarithmic bound on the location of the poles of the scattering operator*, Arch. Rational Mech. Anal., 40 (1971), pp. 268–280.
[14] R.C. MacCamy and E. Stephan, *Solution procedures for three-dimensional eddy current problems*, J. Math. Anal. Appl., 101 (1984), pp. 348–379.
[15] R. Melrose, *Polynomial bounds on the distribution of poles in scattering by an obstacle*, in Journées Equations aux Dérivées Partielles, Saint-Jean-des-Monts, 1984.
[16] R. Melrose, *Geometric scattering theory*, in Stanford Lectures, Cambridge University Press, Cambridge, 1995.
[17] S. Steinberg, *Meromorphic families of compact operators*, Arch. Rational Mech. Anal., 31 (1968), pp. 372–380.
[18] M. Zworski, *Counting scattering poles*, in Spectral and Scattering Theory, M. Dekker, ed., Lecture Notes in Pure and Appl. Math 161, New York, 1994, pp. 301–331.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 879–895

NONUNIQUENESS FOR SECOND-ORDER ELLIPTIC EQUATIONS WITH MEASURABLE COEFFICIENTS*

MIKHAIL V. SAFONOV†

Abstract. We consider the Dirichlet problem for uniformly elliptic operators $L = \sum a_{ij} D_{ij}$ with measurable coefficients a_{ij} in the unit ball $B_1 \subset \boldsymbol{R}^d$. A recent sensational result of Nikolai Nadirashvili states that there is no uniqueness of "weak" solutions to this problem if $d \geq 3$. He constructed two sequences of linear elliptic operators with smooth coefficients $\{a_{ij}^{0,k}\}$ and $\{a_{ij}^{1,k}\}$, which have the same ellipticity constant $\nu > 0$ and converge to the same functions a_{ij} almost everywhere (a.e.) in B_1 as $k \to \infty$, while the corresponding sequences of solutions $\{u^{0,k}\}$ and $\{u^{1,k}\}$ converge to two different functions; i.e., the Dirichlet problem has at least two "weak" solutions. In the present paper, we popularize and slightly generalize Nadirashvili's result: for an arbitrary constant $\Lambda > 0$, we construct two sequences of linear elliptic operators with the same ellipticity constant $\nu = \nu(\Lambda) > 0$ and the additional restriction $|a_{ij}^{0,k} - a_{ij}^{1,k}| \leq \Lambda$ for all i, j, k, which define two different "weak" solutions to the Dirichlet problem [N. S. Nadirashvili, *Ann. Scuola Norm. Sup. Pisa Cl. Sci.* (4), 24 (1997), pp. 537–550].

Key words. elliptic PDE, measurable coefficients, nonuniqueness

AMS subject classifications. 35A05, 35B27, 35J25, 60G44, 60J60

PII. S0036141096309046

1. Introduction.

Let Ω be a bounded domain in $\boldsymbol{R}^d$, $d \geq 1$, with the boundary $\partial\Omega$ of class C^2. Let $a = (a_{ij}(x))$ be a real, symmetric, measurable $d \times d$ matrix function on Ω, satisfying the *uniform ellipticity condition*

$$\nu|\xi|^2 \leq \sum_{i,j=1}^{d} a_{ij}\xi_i\xi_j \leq \nu^{-1}|\xi|^2 \quad \text{for all} \quad \xi \in \boldsymbol{R}^d, \tag{1.1}$$

where $\nu = \text{const} \in (0, 1]$. We consider the *Dirichlet problem*

$$Lu = \sum_{i,j=1}^{d} a_{ij} D_{ij} u = 0 \quad \text{in} \quad \Omega, \qquad u = g \quad \text{on} \quad \partial\Omega, \tag{1.2}$$

where $D_{ij}u = \partial^2 u/\partial x_i \partial x_j$, and g is a given continuous function. If $a_{ij} \in C(\overline{\Omega})$, then this problem has a unique strong solution, which belongs to the Sobolev space $W^{2,p}(\Omega')$ for any subdomain $\Omega' \subset \overline{\Omega'} \subset \Omega$ and $1 < p < \infty$ (see [15, section 9.6]). For discontinuous a_{ij}, there is no definition of solution to the problem (1.2), which preserves simultaneously all the basic properties of strong solutions, such as the existence, uniqueness, maximum principle, etc. For arbitrary approximation a_{ij} by smooth functions a_{ij}^k, $k = 1, 2, \ldots$, the solutions u^k to the corresponding Dirichlet problems are uniformly bounded and equicontinuous on $\overline{\Omega}$ (see [23], [27], and also [15, section 9.8]). Therefore, there exists a subsequence (we call it $\{u^k\}$ again) converging to some function $u \in C(\overline{\Omega})$. In the following definition, this function is called

*Received by the editors September 9, 1996; accepted for publication (in revised form) March 2, 1998; published electronically May 18, 1999. This work was partially supported by National Science Foundation grant DMS-9623287.

http://www.siam.org/journals/sima/30-4/30904.html

†School of Mathematics, University of Minnesota, 127 Vincent Hall, 206 Church Street, Minneapolis, MN 55455 (safonov@math.umn.edu).

a weak solution to the problem (1.2), and so our previous argument shows that the weak solutions always exist.

DEFINITION 1.1. *A function* $u = u(x) \in C(\overline{\Omega})$ *is a* weak solution *to the problem* (1.2) *if there exists a sequence of real symmetric matrices of coefficients* $\{a^k\} = \{(a^k_{ij}(x))\}$, $k = 1, 2, \dots$, *such that*

(i) $a^k_{ij} \in C^\infty(\overline{\Omega})$ *and satisfy* (1.1) *for each* k *with the same constant* $\nu \in (0, 1]$;

(ii) $a^k_{ij} \to a_{ij}$ *a.e. in* Ω *as* $k \to \infty$;

(iii) *the (classical) solutions* $\{u^k\}$ *to the problems*

$$L^k u^k = \sum_{i,j=1}^{d} a^k_{ij} D_{ij} u^k = 0 \quad \text{in} \quad \Omega, \qquad u^k = g \quad \text{on} \quad \partial\Omega \tag{1.3}$$

converge in $C(\overline{\Omega})$ *to* u:

$$\lim_{k\to\infty} \sup_{\Omega} |u^k - u| = 0. \tag{1.4}$$

The properties of classical or strong solutions, which do not depend on the smoothness of coefficients, such as the Hölder estimates, the Harnack inequality, and different forms of the maximum principle, are automatically extended to the weak solutions. As regards the uniqueness of weak solutions, the situation is different. For classical solutions, the uniqueness follows from the maximum principle, because the difference of two solutions to a linear homogeneous equation $Lu = 0$ satisfies the same equation. This reasoning does not work for weak solutions because different weak solutions may be obtained by means of different sequences $\{a^k\}$. Many mathematicians tried to solve this so-called weak uniqueness problem for discontinuous a_{ij} (see [3], [4], [5], [9], [10], [11], [22], [29], and references therein), but positive results were obtained only under additional restrictions on a_{ij}. We will discuss some of these results in section 2 below.

A recent sensational result of Nikolai Nadirashvili [25] states that the weak uniqueness may fail if $d \geq 3$. He constructed two sequences of operators with smooth coefficients $\{a^{0,k}_{ij}\}$ and $\{a^{1,k}_{ij}\}$, which satisfy the ellipticity condition (1.1) with the same constant $\nu \in (0, 1]$ and converge to the same functions a_{ij} a.e. in the unit ball $B_1 = \{|x| < 1\} \subset \boldsymbol{R}^3$ as $k \to \infty$, while the corresponding sequences of solutions converge to two different functions. In other words, there are at least two different weak solutions to the problem (1.2) in $\Omega = B_1 \subset \boldsymbol{R}^3$. In this paper, we show that the weak uniqueness still fails under the additional restriction

$$\sup_{\Omega} |a^{0,k}_{ij} - a^{1,k}_{ij}| \leq \Lambda \qquad \text{for all} \quad i, j, k \tag{1.5}$$

with a constant $\Lambda > 0$ which can be made arbitrarily small. Namely, we have the following result.

THEOREM 1.2. *For any constant* $\Lambda > 0$, *there exist two sequences* $\{a^{0,k}\} = \{(a^{0,k}_{ij})\}$ *and* $\{a^{1,k}\} = \{(a^{1,k}_{ij})\}$ *of real, symmetric,* 3×3 *matrix functions on* $B_1 \subset \boldsymbol{R}^3$, *and a function* $g \in C^\infty(\overline{B_1})$, *such that*

(i) *for each* k, *functions* $\{a^{0,k}_{ij}\}$, $\{a^{1,k}_{ij}\} \subset C^\infty(\overline{B_1})$, *satisfy* (1.5) *with the given constant* Λ, *and* (1.1) *with a constant* $\nu = \nu(\Lambda) \in (0, 1]$;

(ii) $a^{0,k}_{ij}, a^{1,k}_{ij}$ *converge to the same functions* a_{ij} *a.e. in* B_1 *as* $k \to \infty$;

(iii) *for* $m = 0$ *and* 1, *the sequences* $u^{m,k}$ *of solutions to the problems*

$$L^{m,k}u^{m,k} = \sum_{i,j=1}^{3} a_{ij}^{m,k} D_{ij} u^{m,k} = 0 \quad \text{in} \quad B_1, \ u^{m,k} = g \quad \text{on} \quad \partial B_1 \tag{1.6}$$

converge to two different functions u^0 *and* u^1, *respectively:*

$$\lim_{k\to\infty} \sup_{B_1} |u^{m,k} - u^m| = 0, \qquad u^0(0) \neq u^1(0). \tag{1.7}$$

By Definition 1.1, these limit functions u^0 and u^1 are two different weak solutions to the problem (1.2) for $\Omega = B_1$.

Proof. The proof of this theorem is given below in section 4. As in the Nadirashvili paper [25], we use some special "piecewise periodic" approximations of a_{ij}, though the technical details here are different from those in [25].

Theorem 1.2 is easily extended to higher dimensions $d > 3$, with minimal modifications in the proof.

Remark 1.1. The definition of weak solution here and in [22], [29] is equivalent to the definition of "good" solution in [9], [10], [11], [25]. Jensen [16] (see also [8]) proved that the concept of weak solution to (1.2) coincides with that of viscosity solution. Our preference is motivated by the following reasons. By Corollary 2.2 below, any strong solution $u \in W^{2,d}(\Omega)$ of (1.2) is a weak solution, but not vice versa. Moreover, the weak uniqueness problem of (1.2) with arbitrary g is equivalent to the weak uniqueness problem for corresponding diffusion processes (see [19], [21]). This means Nadirashvili's example automatically gives the negative answer to the latter problem.

The following *martingale problem* is also equivalent to the weak uniqueness problem for (1.2) with arbitrary $g \in C(\boldsymbol{R}^d)$ (see [19], [21], and a discussion in [9]). Let an operator $L = \sum a_{ij} D_{ij}$ with coefficients a_{ij} satisfying (1.1) be defined on the whole space $\boldsymbol{R}^d$. Using a "selection" procedure initiated by Krylov [20], Stroock and Varadhan [30] proved that for each $x \in \boldsymbol{R}^d$, there exists a probability measure P_x on $C([0,\infty), \boldsymbol{R}^d) = \{\xi_t, t \geq 0\}$, such that

(i) $P_x(\xi_0 = x) = 1$,

(ii) $\varphi(\xi_t) - \varphi(\xi_0) - \int_0^t L\varphi(\xi_s)\, ds$ is a P_x-local martingale for all $\varphi \in C^2(\boldsymbol{R}^d)$.

They also showed that the solution to the martingale problem is unique if a_{ij} are continuous or $d \leq 2$ (see [31] and [30, Chapter 7]). It remained unknown if the solution was unique for $d \geq 3$ and discontinuous a_{ij}. Nadirashvili's result gives a negative answer to this question as well: the martingale problem may have different solutions.

In section 2, we discuss different conditions on the coefficients, which provide the uniqueness of weak solutions. Section 3 contains some preparatory material. Finally, section 4 is devoted to the proof of Theorem 1.2.

2. Uniqueness of weak solutions. Many positive results on the weak uniqueness are based on the following Aleksandrov–Bakelman–Pucci estimate for functions in the Sobolev space $W^{2,d}(\Omega)$. Having in mind the imbedding $W^{2,d}(\Omega) \subset C(\overline{\Omega})$, we always consider only such "representatives" of functions in $W^{2,d}(\Omega)$ which are continuous on $\overline{\Omega}$.

THEOREM 2.1. *Let* $a = (a_{ij}(x))$ *be a real, symmetric, measurable* $d \times d$ *matrix function on a bounded domain* $\Omega \subset \boldsymbol{R}^d$, $d \geq 2$. *Then for any function* $u \in W^{2,d}(\Omega)$,

we have

$$\sup_{\Omega} |u| \le \sup_{\partial\Omega} |u| + N\|Lu\|_{\mathcal{L}^d(\Omega)}, \tag{2.1}$$

where $L = \sum a_{ij} D_{ij}$*, and the constant* N *depends only on* d*,* ν*, and* $\operatorname{diam}\Omega$*.*

Proof. The proof is contained in [1] (see also [15, Chapter 9]).

COROLLARY 2.2. *If the problem* (1.2) *has a strong solution* $u \in W^{2,d}(\Omega)$*, then* u *is also a unique weak solution to this problem.*

Proof. Let $a^k = (a^k_{ij}(x))$, $k = 1, 2, \dots$, be real, symmetric, smooth matrix functions, such that $a^k_{ij} \to a_{ij}$ a.e. in Ω as $k \to \infty$, and let $\{u^k\}$ be a sequence of solutions to the problems (1.3). Then

$$L^k(u^k - u) = -L^k u = (L - L^k)u \quad \text{in } \Omega, \qquad u^k - u = 0 \text{ on } \partial\Omega.$$

By Theorem 2.1,

$$\begin{aligned} \sup_{\Omega} |u^k - u| &\le N \cdot \|(L - L^k)u\|_{\mathcal{L}^d(\Omega)} \\ &= N \left(\int_\Omega \left| \sum_{i,j} (a_{ij} - a^k_{ij}) D_{ij} u \right|^d dx \right)^{1/d}, \end{aligned} \tag{2.2}$$

with a constant N independent on k. Since the integral function converges to 0 a.e. in Ω as $k \to \infty$ and is dominated by $\text{const} \cdot \sum |D_{ij}u|^d \in \mathcal{L}^1(\Omega)$, we have (1.4). This means u is the only weak solution to the problem (1.2). □

Using suitable approximations of $g \in C(\overline{\Omega})$ by smooth functions, we can reduce the weak uniqueness problem to the case $g \in C^\infty(\overline{\Omega})$. Further, we set $v = u - g$, so that the problem (1.2) is transformed to the equivalent one:

$$Lv = \sum_{i,j} a_{ij} D_{ij} v = f \quad \text{in} \quad \Omega, \qquad v = 0 \quad \text{on} \quad \partial\Omega, \tag{2.3}$$

where $f = -Lg \in \mathcal{L}^\infty(\Omega)$. The problem (2.3) has a unique strong solution in $W^{2,d}(\Omega)$, even for $f \in \mathcal{L}^d(\Omega)$, if the estimate

$$\max_{i,j} \|D_{ij} v\|_{\mathcal{L}^d(\Omega)} \le N \|Lv\|_{\mathcal{L}^d(\Omega)} \quad \text{for} \quad v \in C^2(\overline{\Omega}), \quad v = 0 \quad \text{on} \quad \partial\Omega, \tag{2.4}$$

holds with a constant N independent of v (see [15, Chapter 9]). This is true in the following cases (i), (ii), and (iii).

(i) $a_{ij} \in C(\overline{\Omega})$. By the Calderon–Zygmund inequality (see [15, Chapter 9]), the estimate (2.4) remains true also for norms in $\mathcal{L}^p(\Omega)$, $1 < p < \infty$.

(ii) $d = 2$. The estimate (2.4) is due to S. N. Bernstein when Ω is a ball in $\boldsymbol{R}^2$. It is generalized to bounded domains Ω with $\partial\Omega \in C^2$ (see [24, Chapter 3, section 19]).

(iii) The coefficients a_{ij} are close to some constants a^0_{ij} (cf. [12]):

$$|a_{ij} - a^0_{ij}| \le \varepsilon_0, \tag{2.5}$$

where $\varepsilon_0 = \varepsilon_0(d, \nu)$ is a small positive constant. The estimate (2.4) for $L = \sum a_{ij} D_{ij}$ is obtained by easy perturbation arguments. Indeed, as a particular case of (i), this

estimate holds for $L^0 = \sum a^0_{ij} D_{ij}$; i.e., for arbitrary $v \in C^2(\overline{\Omega})$ satisfying $v = 0$ on $\partial\Omega$, we have

$$M = \sum_{i,j} \|D_{ij} v\| \le N_0 \|L^0 v\|,$$

where $N_0 = N_0(\nu, \Omega)$ and $\|\cdot\|$ is the norm in $\mathcal{L}^d(\Omega)$. Using (2.5), we get

$$\|L^0 v\| \le \|Lv\| + \|(L - L^0)v\| \le \|Lv\| + \varepsilon_0 M,$$

$$M \le N_0 \|L^0 v\| \le N_0 \|Lv\| + N_0 \varepsilon_0 M.$$

For $0 < \varepsilon_0 < 1/N_0$, the last inequality implies (2.4) with $N = (1 - N_0\varepsilon_0)^{-1} N_0$.

Summarizing the above considerations, we obtain the weak uniqueness for the problem (1.2) with arbitrary $g \in C(\overline{\Omega})$ in the cases (i), (ii), and (iii).

Corollary 2.2 does not cover all the cases when the weak uniqueness holds. If the coefficients a_{ij} are discontinuous at a single point, the weak solution to the problem (1.2) may not belong to $W^{2,p}(\Omega)$ if $p > \frac{3}{2}$ (see [28, Remark 8.1]). However, Luis Caffarelli proved that in this case the weak solution is unique. Relying on this result, Cerutti, Escauriaza, and Fabes [9], [10] proved the uniqueness when $a_{ij} \in C(\Omega \setminus E)$, where E is a countable set having at most one cluster point. Further generalizations were made by Krylov [22] for E with countable closure and by Safonov [29] for closed E of small Hausdorff dimension $\alpha = \alpha(d, \nu)$; i.e., $\alpha \to 0$ as $\nu \to 0$. Nevertheless, many special questions remain open, e.g., whether or not the weak uniqueness holds when E is a segment in $\Omega \subset \boldsymbol{R}^d$, $d \ge 3$.

For other related results, see [3], [5], [4], [11]. We only mention a paper by Bass and Pardoux [5], where the weak uniqueness is proved when Ω is the union of a finite number of disjoint polyhedrons and a_{ij} are constants on each of them.

Since the problems (1.2) and (2.3) are equivalent, the weak uniqueness can be treated in terms of properties of the Green function for Ω (see [9], [10], [11], [22]). By this approach, Krylov [22] showed that the weak uniqueness is a local property of coefficients a_{ij}, i.e., it holds for a neighborhood of each point $x_0 \in \Omega$, if and only if it holds for Ω.

Both (1.2) and (2.3) are the particular cases of the problem

$$Lu = \sum_{i,j} a_{ij} D_{ij} u = f \quad \text{in} \quad \Omega, \qquad u = g \quad \text{on} \quad \partial\Omega, \tag{2.6}$$

where $f \in \mathcal{L}^d(\Omega)$, $g \in C(\overline{\Omega})$. The weak uniqueness for (1.2) with arbitrary $g \in C(\overline{\Omega})$ implies the uniqueness for (2.6) in the following sense. Let smooth functions $a^k_{ij} = a^k_{ji}$, f^k, g^k be given for $i, j = 1, 2, \dots, d$; $k = 1, 2, \dots$. Suppose a^k_{ij} satisfy (1.1) for each k, $a^k_{ij} \to a_{ij}$ a.e. in Ω, $f^k \to f$ in $\mathcal{L}^d(\Omega)$, and $g^k \to g$ in $C(\overline{\Omega})$ as $k \to \infty$. Let u^k be a (classical) solution to the problem

$$L^k u^k = \sum_{i,j} a^k_{ij} D_{ij} u^k = f^k \quad \text{in} \quad \Omega, \qquad u^k = g^k \quad \text{on} \quad \partial\Omega. \tag{2.7}$$

Then $u^k \to u$ in $C(\overline{\Omega})$ as $k \to \infty$, where the limit function u does not depend on the choice of a^k_{ij}, f^k, g^k (see [29, section 2]). The weak uniqueness is also preserved after perturbation of L by zero order terms (see [22], [11]). However, this is not true for

approximations of f in $\mathcal{L}^p(\Omega)$ with $p < d$, as the following example shows (cf. [14], [2], [26]). This example also shows that the estimate (2.1) fails for $\mathcal{L}^p$-norms, $p < d$, in place of the $\mathcal{L}^d$-norm.

Example 2.1. Let $d \geq 2$, $0 < p < d$ be fixed. We choose $\alpha = \text{const} \in (0,1)$, such that $(2-\alpha)p < d$, and consider the function $u(x) = 1 - |x|^\alpha$ on $B_1 \subset \boldsymbol{R}^d$. Since $D_{ij}u(x) = O(|x|^{\alpha-2})$, the function u belongs to $W^{2,p}(B_1)$ and satisfies

$$Lu = \sum_{i,j} a_{ij} D_{ij} u = 0 \quad \text{in} \quad B_1, \qquad u = 0 \quad \text{on} \quad \partial B_1, \tag{2.8}$$

where

$$a_{ij} = \delta_{ij} + \beta \frac{x_i x_j}{|x|^2}, \qquad \beta = \frac{d-2+\alpha}{1-\alpha}.$$

Obviously, the only weak solution to (2.8) is identically zero. Although our function $u(x)$ has a singularity at 0, we can approximate it by functions u^ε, $0 < \varepsilon < 1$, such that

$$u^\varepsilon \in C^\infty(\overline{\Omega}), \quad u^\varepsilon = u \quad \text{on} \quad B_1 \setminus B_\varepsilon; \qquad |D_{ij}u^\varepsilon| \leq N\varepsilon^{\alpha-2} \quad \text{on} \quad B_\varepsilon,$$

with a constant N which is independent on ε. Therefore,

$$Lu^\varepsilon = f^\varepsilon \quad \text{in} \quad B_1, \qquad u^\varepsilon = 0 \quad \text{on} \quad \partial B_1,$$

where

$$f^\varepsilon = 0 \quad \text{on} \quad B_1 \setminus B_\varepsilon, \qquad |f^\varepsilon| \leq N\varepsilon^{\alpha-2} \quad \text{on} \quad B_\varepsilon.$$

In particular, $f^\varepsilon \to 0$ in $\mathcal{L}^p(B_1)$, while $u^\varepsilon \to u \neq 0$ in $C(\overline{B_1})$ as $\varepsilon \to 0$.

3. Auxiliary results. In this section, we consider linear elliptic operators $L = \sum a_{ij} D_{ij}$ with the coefficients $a_{ij} \in C^\infty(\boldsymbol{R}^d)$, $d \geq 1$. As usual, we assume that the matrix function $a = (a_{ij}(x))$ is real, symmetric, and satisfies (1.1) with a constant $\nu \in (0,1]$. In addition, suppose a_{ij} are periodic with period 1 in each of the variables. We may treat a_{ij} as functions in $C^\infty(\boldsymbol{T}^d)$, where $\boldsymbol{T}^d$ is the d-dimensional torus. We will identify $\boldsymbol{T}^d$ with $[-\frac{1}{2}, \frac{1}{2})^d \subset \boldsymbol{R}^d$ and the functions on $\boldsymbol{T}^d$ with 1-periodic functions on $\boldsymbol{R}^d$, so that

$$u(x) \equiv u(x+z) \quad \text{on} \quad \boldsymbol{T}^d = \left[\frac{1}{2}, \frac{1}{2}\right)^d \tag{3.1}$$

for some vector z with integer components.

The following theorem is a version of the Fredholm alternative for equations with periodic coefficients.

THEOREM 3.1.

(i) *The equation*

$$Lu = \sum_{i,j} a_{ij} D_{ij} u = f \tag{3.2}$$

with $f \in C^\infty(\boldsymbol{T}^d)$ has a solution in $C^\infty(\boldsymbol{T}^d)$ if and only if

$$(f, \rho) = \int_{\boldsymbol{T}^d} f\rho \, dx = 0 \tag{3.3}$$

for every solution $\rho \in C^\infty(\boldsymbol{T}^d)$ *of the adjoint homogeneous equation*

$$L^*\rho = \sum_{i,j} D_{ij}(a_{ij}\rho) = 0. \tag{3.4}$$

(ii) *The homogeneous equations* $Lu = 0$ *and* $L^*\rho = 0$ *have the same finite number of linearly independent solutions.*

Proof. The proof is given in [7, part 2, section 3.6], even for more general equations of higher order. For second-order equations, we specify these statements as follows.

COROLLARY 3.2. *Under the additional restriction*

$$\int_{\boldsymbol{T}^d} \rho\, dx = 1, \tag{3.5}$$

equation (3.4) *has a unique solution* $\rho \in C^\infty(\boldsymbol{T}^d)$. *Moreover,* $\rho \geq 0$ *on* $\boldsymbol{T}^d$. *Equation* (3.2) *with* $f \in C^\infty(\boldsymbol{T}^d)$ *has a solution in* $C^\infty(\boldsymbol{T}^d)$ *if and only if* f *satisfies* (3.3) *with this* ρ.

Proof. By the strong maximum principle (see [7, part 2, section 2.2], or [15, Theorem 3.5]), from $Lu = f \geq 0$ on $\boldsymbol{T}^d$ it follows that $u = \text{const}$, $f = 0$. Therefore, the number of linearly independent solutions for each of two equations $Lu = 0$ and $L^*\rho = 0$ is one. This implies that the problem (3.4), (3.5) has a unique solution $\rho \in C^\infty(\boldsymbol{T}^d)$, and the last statement here is equivalent to Theorem 3.1(i).

To prove the inequality $\rho \geq 0$ on $\boldsymbol{T}^d$, suppose otherwise. Then there exists a positive function $f \in C^\infty(\boldsymbol{T}^d)$ satisfying (3.3) with this ρ. In turn, this provides the solvability of (3.2) for $f > 0$, which is impossible. This contradiction gives $\rho \geq 0$, and the proof is completed. □

DEFINITION 3.3. *Let* $a = (a_{ij})$ *be a real, symmetric* $d \times d$ *matrix function, where* $a_{ij} \in C^\infty(\boldsymbol{T}^d)$ *and satisfy* (1.1) *with a constant* $\nu \in (0, 1]$. *A* homogenized matrix *of* a *is the constant matrix*

$$a^0 = (a^0_{ij}) = H(a) = \int_{\boldsymbol{T}^d} a\rho\, dx, \tag{3.6}$$

where $\rho \in C^\infty(\boldsymbol{T}^d)$ *is a (unique) solution to the problem* (3.4), (3.5).

The following theorem was proved by Freidlin [13]. He used an interpretation of $\rho(x)$ as the density of limit distributions on $\boldsymbol{T}^d$ of Markov processes corresponding to $L = \sum a_{ij}D_{ij}$. Here we give an alternative proof which is closer to [6], [17]. We do not use this theorem here; however, its proof presents in a "pure" form the basic element in the proof of Theorem 1.2.

THEOREM 3.4. *Let* $d \times d$ *matrices* $a(x)$ *and* a^0 *be as in Definition* 3.3. *For* $\varepsilon > 0$, *we set* $a^\varepsilon(x) = a(\varepsilon^{-1}x)$, *so that* a^ε *are defined for all* $\varepsilon \geq 0$. *Let* Ω *be a bounded domain in* $\boldsymbol{R}^d$ *with the boundary* $\partial\Omega \in C^\infty$, *and let a function* $g \in C^\infty(\overline{\Omega})$ *be given. Then the solutions* u^ε, $\varepsilon \geq 0$, *to the problems*

$$L^\varepsilon u^\varepsilon = \sum_{i,j} a^\varepsilon_{ij} D_{ij} u^\varepsilon = 0 \quad \text{in } \Omega, \quad u^\varepsilon = g \quad \text{on } \partial\Omega, \tag{3.7}$$

satisfy

$$\lim_{\varepsilon \to 0+} \sup_{\Omega} |u^\varepsilon - u^0| = 0. \tag{3.8}$$

Proof. We have the matrix equality

$$(a-a^0,\rho)=\int_{\boldsymbol{T}^d}(a-a^0)\rho\,dx=\int_{\boldsymbol{T}^d}a\rho\,dx-a^0\int_{\boldsymbol{T}^d}\rho\,dx=a^0-a^0=0.$$

By Corollary 3.2 applied to each entry of the matrix $a-a^0$, there exists a matrix function $v=(v_{ij})$ with $v_{ij}\in C^\infty(\overline{\Omega})$, such that

$$Lv(x)=\sum_{i,j}a_{ij}D_{ij}v(x)=a(x)-a^0. \tag{3.9}$$

For $\varepsilon>0$, we have

$$\varepsilon^2L^\varepsilon(v(\varepsilon^{-1}x))=(L^\varepsilon v)(\varepsilon^{-1}x)=a(\varepsilon^{-1}x)-a^0=a^\varepsilon(x)-a^0. \tag{3.10}$$

Consider the functions

$$w^\varepsilon(x)=u^\varepsilon(x)-u^0(x)+\varepsilon^2\sum_{i,j}v_{ij}(\varepsilon^{-1}x)D_{ij}u^0(x).$$

Since the coefficients and the boundary data in (3.7) are smooth, all the functions u^ε, u^0, v_{ij} belong to $C^\infty(\overline{\Omega})$. Using the identities $L^\varepsilon u^\varepsilon=L^0u^0=0$ and (3.10), we get

$$L^\varepsilon w^\varepsilon=L^\varepsilon u^\varepsilon-L^\varepsilon u^0+\varepsilon^2\sum_{i,j}L^\varepsilon(v_{ij}(\varepsilon^{-1}x))\cdot D_{ij}u^0(x)+F^\varepsilon$$

$$=L^\varepsilon u^\varepsilon-L^\varepsilon u^0+(L^\varepsilon-L^0)u^0+F^\varepsilon=F^\varepsilon,$$

where

$$F^\varepsilon(x)=\varepsilon^2\sum_{i,j}\left[L^\varepsilon(v_{ij}(\varepsilon^{-1}x)D_{ij}u^0(x))-L^\varepsilon(v_{ij}(\varepsilon^{-1}x))\cdot D_{ij}u^0(x)\right]$$

satisfies $|F^\varepsilon|\le N\varepsilon$ with a constant N which is independent of ε. Moreover, since $u^\varepsilon=u^0=g$ on $\partial\Omega$, we also have $|w^\varepsilon|\le N\varepsilon^2$ on $\partial\Omega$. By the comparison principle for second-order elliptic equations, $|w^\varepsilon|\le N\varepsilon$ in Ω for $0<\varepsilon<1$, which gives us equation (3.8). □

As in the Nadirashvili paper [25], our approach to Theorem 1.2 is based on the fact that two different constant 3×3 matrices A^0 and A^1 may be obtained by homogenization of smooth matrices a^0 and a^1 correspondingly, which coincide on a portion of $\boldsymbol{T}^3$, and on the fact that locally the structure of a^m is similar to that of A^m ($m=0$ and 1) on the remaining part of $\boldsymbol{T}^3$. The details are given in the following technical lemma.

LEMMA 3.5. *For any constants $\Lambda\ge\lambda_0>0$ and $\theta>0$, there exists an open set $G=G(\Lambda,\theta)\subset\boldsymbol{T}^3$ of the Lebesgue measure $|G|\le\theta$ and two real symmetric, 3×3 matrix functions $a^0=(a^0_{ij}(x))$, $a^1=(a^1_{ij}(x))$, such that*

(i) *functions $a^0_{ij},a^1_{ij}\in C^\infty(\boldsymbol{T}^3)$, satisfy* (1.1) *with a constant $\nu=\nu(\Lambda,\theta)\in(0,1]$, and*

$$a^0=I,\quad a^1=I+\lambda\,l^tl\quad\text{on}\quad G,\qquad a^0=a^1\quad\text{on}\quad\boldsymbol{T}^3\setminus G, \tag{3.11}$$

where $I=(\delta_{ij})$ is the unit matrix, $\lambda\in C^\infty(\boldsymbol{T}^3)$, $l=(l_1,l_2,l_3)$ with $l_i\in C^\infty(G)$, l^t is the corresponding column vector, and

$$0\le\lambda\le\Lambda,\ \ |l|^2=\sum_i l_i^2=1\quad\text{on}\quad G,\qquad\lambda\equiv0\quad\text{on}\quad\boldsymbol{T}^3\setminus G; \tag{3.12}$$

(ii) *the homogenized matrices*

$$A^0 = H(a^0) = I, \qquad A^1 = H(a^1) = c(I + \lambda_0 e_3^t e_3), \tag{3.13}$$

where $1 \le c = \text{const} \le 1 + \Lambda$, $e_3 = (0, 0, 1)$.

Proof. Step 1. We will find G in the form $G = G' \times \boldsymbol{T}^1 \subset \boldsymbol{T}^3$, where

$$G' = \{x' = (x_1, x_2) : r_0 < |x'| < r_1\} \subset \boldsymbol{T}^2 = \left[-\frac{1}{2}, \frac{1}{2}\right)^2, \tag{3.14}$$

$r_1 = \min(\frac{1}{4}, \sqrt{\frac{\theta}{\pi}})$, and a small constant $r_0 \in (0, \frac{1}{4} r_1]$ will be chosen later. We have $|G| = |G'| < \pi r_1^2 \le \theta$. Next, we choose a smooth function λ on $\boldsymbol{R}^1$ with compact support in (r_0, r_1) such that

$$0 \le \lambda \le \Lambda \quad \text{on } \boldsymbol{R}^1, \qquad \lambda = \Lambda \quad \text{on } \left[2r_0, \frac{1}{2} r_1\right]. \tag{3.15}$$

We will use the same notation λ for the function $\lambda = \lambda(r)$, $r = |x'| = \sqrt{x_1^2 + x_2^2}$ on $\boldsymbol{R}^2 \subset \boldsymbol{R}^3$. Further, we set

$$l = (l_1, l_2, l_3) = r^{-1}(x_1, x_2, 0), \quad r = |x'| \quad \text{on} \quad G. \tag{3.16}$$

Then λ and l satisfy (3.12), and the matrices a^0, a^1 are well defined on G by the identities (3.11). We will extend $a^0_{33} = a^1_{33} = 1$ from G to $\boldsymbol{T}^3$ by the formula

$$a^0_{33} = a^1_{33} = 1 + c_0 \zeta_0(r) - c_1 \zeta_1(r), \qquad r = |x'| \tag{3.17}$$

with some constants $c_0 \ge 0$, $0 \le c_1 \le \frac{1}{2}$. Here ζ_0 and ζ_1 are smooth functions on $\boldsymbol{R}^1$ with compact supports in $(-r_0, r_0)$ and $(\frac{1}{4}, \frac{1}{2})$, correspondingly, satisfying $0 \le \zeta_0, \zeta_1 \le 1$ on $\boldsymbol{R}^1$, $\zeta_0 = 1$ on $[-\frac{1}{2} r_0, \frac{1}{2} r_0]$, and $\zeta_1 = 1$ on a fixed subinterval of $(\frac{1}{4}, \frac{1}{2})$. The remaining coefficients

$$a^0_{ij} = a^1_{ij} = \delta_{ij} \quad \text{on} \quad \boldsymbol{T}^3 \setminus G, \quad i + j < 6. \tag{3.18}$$

Together with (3.11), the formulas (3.17), (3.18) determine the matrices $a^0, a^1 \in C^\infty(\boldsymbol{T}^3)$ satisfying all the properties (i). The properties (ii) will be obtained by a suitable choice of the constants r_0, c_0, c_1.

Step 2. For $m = 0$ and 1, we denote by ρ_m the solution to the problem (3.4), (3.5) with $d = 3$, $a_{ij} = a^m_{ij}$. Since $a^m = (a^m_{ij})$ depend only on $x' = (x_1, x_2)$, ρ_m coincide with the (unique) solutions to the problem

$$(L^m)^* \rho_m = \sum_{i,j=1}^{2} D_{ij}(a^m_{ij} \rho_m(x')) = 0, \qquad \int_{\boldsymbol{T}^2} \rho_m(x')\, dx' = 1. \tag{3.19}$$

Therefore, $\rho_m(x) = \rho_m(x') = \rho_m(x_1, x_2)$ do not depend on x_3. By (3.11) and (3.18), $a^0_{ij} = \delta_{ij}$ for $i + j < 6$. Hence $\rho_0 \equiv 1$, and the homogenized matrix $A^0 = H(a^0)$ has the entries

$$A^0_{ij} = \int_{\boldsymbol{T}^2} a^0_{ij}\, dx' = \delta_{ij}, \quad i + j < 6.$$

We do not have any explicit representations for ρ_1. However, from the symmetry it follows that $\rho_1 = \rho_1(x_1, x_2)$ is an even function of x_1 and x_2. By (3.11), (3.15)–(3.18), we get

$$A^1_{ij} = \int_{\boldsymbol{T}^2} a^1_{ij}\rho_1\,dx' = \delta_{ij} + \int_{\boldsymbol{T}^2} \lambda r^{-2}x_ix_j\rho_1\,dx' = 0, \qquad i \neq j,$$

$$1 \le A^1_{11} = A^1_{22} = 1 + \int_{\boldsymbol{T}^2} \lambda r^{-2}x_1^2\rho_1\,dx' \le 1 + \Lambda,$$

so that

$$A^1_{ij} = c\delta_{ij} \qquad \text{for} \quad i + j < 6, \qquad \text{where} \quad 1 \le c \le 1 + \Lambda. \tag{3.20}$$

Now it remains to get the equalities $A^0_{33} = 1$, $A^1_{33} = c(1 + \lambda_0)$, which by virtue of (3.17) are equivalent to the system

$$\rho_{00}c_0 - \rho_{01}c_1 = 0, \qquad \rho_{10}c_0 - \rho_{11}c_1 = b,$$

where

$$0 \le \rho_{mn} = \int_{\boldsymbol{T}^2} \zeta_n\rho_m\,dx', \qquad 0 \le b = (1 + \lambda_0)c - 1 \le (2 + \Lambda)\Lambda.$$

The solution of this system is

$$c_1 = \frac{b}{K\rho_{01} - \rho_{11}}, \quad c_0 = \frac{\rho_{01}}{\rho_{00}}c_1, \quad \text{where} \quad K = \frac{\rho_{10}}{\rho_{00}}.$$

We also need the inequalities $0 \le c_1 \le \frac{1}{2}$, which guarantee the uniform ellipticity of the matrices a^0, a^1. We may assume that ζ_1 is fixed and for different $r_0 > 0$ the functions ζ_0 are obtained by rescaling $x \longrightarrow \text{const} \cdot x$, so that now everything depends only on $r_0 \in (0, \frac{1}{4}r_0]$. Since ρ_{01} does not depend on r_0 and $\rho_{11} \le 1$, the desired inequalities $0 \le c_1 \le \frac{1}{2}$ hold automatically for small $r_0 > 0$ if we show that

$$K = \frac{\rho_{10}}{\rho_{00}} \to +\infty \quad \text{as} \quad r_0 \to 0+. \tag{3.21}$$

Step 3. We set $B = B_A = \{y' = (y_1, y_2) : |y'| < A\} \subset \boldsymbol{R}^2$, where $A = \frac{\sqrt{2}}{2}$. Notice that for each $x' = (x_1, x_2) \in \boldsymbol{T}^2 = [-\frac{1}{2}, \frac{1}{2})^2$, there exist at least one and at most two points $y' = (y_1, y_2) \in B$, such that both $x_1 - y_1$ and $x_2 - y_2$ are integers. Therefore, considering $\rho_m \in C^\infty(\boldsymbol{T}^2)$ as periodic functions on $\boldsymbol{R}^2$ for $m = 0$ and 1, we obtain

$$1 = \int_{\boldsymbol{T}^2} \rho_m\,dx' \le \int_B \rho_m\,dx' \le 2. \tag{3.22}$$

Further, since $\lambda(r) = 0$, $\zeta_0(r) = 1$ on $[0, \frac{1}{2}r_0]$, the ordinary differential equation

$$(1 + m\lambda)w_m'' + \frac{1}{r}w_m' = \zeta_0 \tag{3.23}$$

has a solution $w_m(r) = \frac{1}{4}r^2$ on $(0, \frac{1}{2}r_0]$. This solution is uniquely extended to $(0, A]$. The function

$$W_m(r) = w_m(r) - \frac{w_m'(A)}{2A}r^2, \quad \text{where} \quad r = |x'| = \sqrt{x_1^2 + x_2^2},$$

belongs to $C^\infty(\overline{B})$ and satisfies

$$L^m W_m = \sum_{i,j=1}^{2} a_{ij}^m D_{ij} W_m = (1+m\lambda)W_m'' + \frac{1}{r}W_m' = \zeta_0 - \frac{w_m'(A)}{A}(2+m\lambda) \tag{3.24}$$

in B, $D_i W_m = 0$ on ∂B. Integrating by parts twice, using (3.19) and then (3.24), we have

$$\int_B L^m W_m \cdot \rho_m \, dx' = \int_B L^m (W_m - W_m(A)) \cdot \rho_m \, dx'$$

$$= \int_B (W_m - W_m(A)) \cdot (L^m)^* \rho_m \, dx' = 0,$$

$$\rho_{m0} = \int_{\boldsymbol{T}^2} \zeta_0 \rho_m \, dx' = \int_B \zeta_0 \rho_m \, dx' = \frac{w_m'(A)}{A} \int_B (2+m\lambda)\rho_m \, dx'.$$

By the inequalities (3.22), we get

$$K = \frac{\rho_{10}}{\rho_{00}} \geq \frac{1}{2} f(A), \qquad \text{where} \quad f(r) = \frac{w_1'(r)}{w_0'(r)}. \tag{3.25}$$

Since $\lambda = 0$ on $[0, r_0]$, we have $w_1 \equiv w_0$ on $[0, r_0]$; in particular, $f(r_0) = 1$. Moreover, since $\zeta_0 = 0$ on $[r_0, \infty)$, and $\lambda = \Lambda$ on $[2r_0, \frac{1}{2}r_1]$, we also have

$$(\ln f)' = \frac{w_1''}{w_1'} - \frac{w_0''}{w_0'} = \frac{-1}{(1+\lambda)r} + \frac{1}{r} = \frac{\lambda}{(1+\lambda)r} \quad \text{on} \quad [r_0, \infty),$$

$$\ln f(A) = \int_{r_0}^{A} \frac{\lambda \, dr}{(1+\lambda)r} \geq \frac{\Lambda}{1+\Lambda} \int_{2r_0}^{\frac{1}{2}r_1} \frac{dr}{r} = \frac{\Lambda}{1+\Lambda} \ln \frac{r_1}{4r_0}.$$

The last relation shows that by the choice of small $r_0 \in (0, \frac{1}{4}r_1]$, $f(A)$ can be made arbitrarily large. Then (3.25) gives us the desired estimate (3.24), and so Lemma 3.5 is proved. □

4. Proof of Theorem 1.2. *Step* 1. For the given constant $\Lambda > 0$ and $\theta = \frac{1}{2}$, we fix the open set $G \subset \boldsymbol{T}^3$ and the constant $\nu = \nu(\Lambda) \in (0, 1]$ from Lemma 3.5. Starting from the unit ball B_1 in $\boldsymbol{R}^3$, we will construct a decreasing sequence of open sets

$$B_1 = G_0 \supset G_1 \supset \cdots \supset G_k \supset G_{k+1} \supset \cdots \tag{4.1}$$

of the Lebesgue measures

$$|G_k| \leq 2^{-k} |B_1|, \qquad k = 0, 1, \ldots, \tag{4.2}$$

and two sequences of real, symmetric, smooth 3×3 matrix functions $\{a^{0,k}\}$ and $\{a^{1,k}\}$ on $\overline{B_1}$, satisfying (1.1) with the chosen constant $\nu = \nu(\Lambda)$, and such that

$$a^{0,k} = I, \quad a^{1,k} = I + \lambda_k (l^k)^t l^k \ \text{ on } \ G_k; \quad a^{0,k} = a^{1,k} = a \ \text{ on } \ B_1 \setminus G_k, \tag{4.3}$$

where $\lambda_k \in C^\infty(\overline{B_1})$, $l^k = (l_1^k, l_2^k, l_3^k)$ with $l_i^k \in C^\infty(\overline{G_k})$, and

$$0 \le \lambda_k \le \Lambda, \quad |l^k| = 1 \quad \text{on} \quad G_k; \qquad \lambda_k \equiv 0 \qquad \text{on} \quad B_1 \setminus G_k. \tag{4.4}$$

Our goal is to construct $\{a^{0,k}\}$ and $\{a^{1,k}\}$ in such a manner that for some $g \in C^\infty(\overline{B_1})$, the sequences $\{u^{0,k}\}$ and $\{u^{1,k}\}$ of solutions to the problems (1.6) uniformly converge to two different functions u^0 and u^1.

Notice that from (4.2)–(4.3) the convergence follows of $a_{ij}^{0,k}$, $a_{ij}^{1,k}$ to a_{ij} a.e. on B_1 as $k \to \infty$; i.e., see statement (ii) of Theorem 1.2.

We set $G_0 = B_1$, $\lambda_0 = \Lambda$, $l^0 = e_3 = (0,0,1)$, $g = x_1^2 - x_3^2$. Then $a^{0,0} = I$, $a^{1,0} = I + \Lambda e_3^t e_3$, the corresponding solutions to the problems (1.6), $u^{0,0} \equiv g$, and $u^{1,0}$ are different; hence

$$|u^{0,0} - u^{1,0}|_0 \ge H \tag{4.5}$$

with some constant $H > 0$. Here and in the remaining part of the proof, $|\cdot|_0$ denotes the norm in $C(\overline{B_1})$.

Step 2. Now we have G_k, $a^{0,k}$, $a^{1,k}$ satisfying (4.2)–(4.4) for $k = 0$. Assuming that they are given for some integer $k \ge 0$, we will construct G_{k+1}, $a^{0,k+1}$, $a^{1,k+1}$. By induction, we get all of these objects for all integers $k \ge 0$.

We divide G_k into a finite number of regular sets $G_{k,n}$, $n = 1, 2, \dots$, fix an arbitrary point $P_{k,n} \in G_{k,n}$ for each n, and introduce the piecewise constant matrix function on G_k,

$$\overline{a} = (\overline{a}_{ij}) = I + \overline{\lambda}\, \overline{l}^t \overline{l}, \quad \text{where} \quad \overline{\lambda} = \lambda_k(P_{k,n}), \ \ \overline{l} = l^k(P_{k,n}) \quad \text{on} \quad G_{k,n}. \tag{4.6}$$

For arbitrary constant $\delta_k > 0$, we can choose $G_{k,n}$ of appropriately small diameters, such that

$$|a_{ij}^{1,k} - \overline{a}_{ij}| \le \delta_k \qquad \text{on} \quad G_k = \bigcup_n G_{k,n}. \tag{4.7}$$

Further, for arbitrary constant $\mu_k > 0$, there exists a function $\eta_k \in C^\infty(\overline{B_1})$, such that $0 \le \eta_k \le 1$ on B_1, $\eta_k = 0$ on $B_1 \setminus G_k$ and near $\partial G_{k,n}$ for each n, and

$$|G_k \setminus G_k'| \le \mu_k, \quad \text{where} \quad G_k' = \{x \in G_k : \eta_k(x) = 1\}. \tag{4.8}$$

The constants δ_k, μ_k will be selected later.

Next, we concentrate on $G_{k,n}$ for fixed n. We change the coordinates by the formula

$$x = yQ_n, \qquad y = xQ_n^t, \tag{4.9}$$

where Q_n is a matrix of rotation satisfying $\overline{l} = e_3 Q_n$. Let $\tilde{a}^0(y)$, $\tilde{a}^1(y)$ be matrix functions a^0, a^1 in Lemma 3.5 applied with $\lambda_0 = \overline{\lambda}$. By (3.13), we have

$$\tilde{A}^0 = H(\tilde{a}^0) = I, \quad \tilde{A}^1 = H(\tilde{a}^1) = c(I + \overline{\lambda} e_3^t e_3), \quad 1 \le c \le 1 + \Lambda.$$

As in (3.9), there exist 3×3 matrix functions $\tilde{v}^m(y) = (\tilde{v}_{ij}^m(y))$ for $m = 0$ and 1, with entries in $C^\infty(\boldsymbol{T}^3)$, satisfying

$$\sum_{i,j} \tilde{a}_{ij}^m(y) D_{ij} \tilde{v}^m(y) = \tilde{a}^m(y) - \tilde{A}^m.$$

One can see from the proof of Lemma 3.5 that the smoothness of $\tilde{a}^m$ does not depend on n (i.e., on the choice of $\lambda_0 = \overline{\lambda}$). Moreover, replacing $\tilde{v}^m$ by $\tilde{v}^m + \text{const}$, we may assume $\tilde{v}^m(0) = 0$. Then by the well-known properties of solutions of elliptic equations with smooth coefficients, the bounds for $\tilde{v}^m$ and their derivative of any order do not depend on n. In the x-coordinates, the matrices $\tilde{a}^m(y)$, $\tilde{v}^m(y)$, $\tilde{A}^m$ are replaced by

$$a^m(x) = Q_n^t \tilde{a}^m(xQ_n^t)Q_n, \quad v^m(x) = Q_n^t \tilde{v}^m(xQ_n^t)Q_n, \quad A^m(x) = Q_n^t \tilde{A}^m Q_n,$$

correspondingly. As a result, we obtain

$$\sum_{i,j} a_{ij}^m D_{ij} v^m(x) = a^m(x) - A^m. \tag{4.10}$$

By (4.6) and the choice of the orthogonal matrix Q_n in (4.9),

$$A^0 = I, \quad A^1 = c(I + \overline{\lambda}\,\overline{l}^t\overline{l}) = c\,\overline{a}, \quad 1 \le c \le 1 + \Lambda. \tag{4.11}$$

Step 3. For $m = 0$ and 1, we define

$$a^{m,k+1}(x) = (1 - \eta_k(x))a^{0,k}(x) + \eta_k(x)\, a^m(\varepsilon^{-1}x) \quad \text{on} \quad B_1 \tag{4.12}$$

with a small $\varepsilon > 0$. Here and in (4.10), the matrix functions a^m, v^m depend also on k, n: these functions satisfy (4.10), (4.11) on $G_{k,n}$ with $\overline{a}$ depending on k, n, so that

$$\varepsilon^2 \sum_{i,j} a_{ij}^m(\varepsilon^{-1}x) D_{ij} v^m(\varepsilon^{-1}x) = a^m(\varepsilon^{-1}x) - A^m, \tag{4.13}$$

but they may be discontinuous on $G_k = \bigcup_n G_{k,n}$. However, since η_k vanishes on $B_1 \setminus G_k$ and near $\partial G_{k,n}$, the functions $a^{0,k+1}$ and $a^{1,k+1}$ are well defined, belong to $C^\infty(\overline{B_1})$, and satisfy (1.1) with the constant $\nu = \nu(\Lambda)$.

Further, we denote by Z^3 the set of all vectors in $\boldsymbol{R}^3$ with integer components and define

$$\boldsymbol{T}_{\varepsilon,n}(z) = \left\{ x \in \boldsymbol{R}^3 : \varepsilon^{-1} x Q_n^t - z \in \boldsymbol{T}^3 = \left[-\frac{1}{2}, \frac{1}{2}\right)^3 \right\},$$

$$G_{\varepsilon,n}(z) = \{x \in \boldsymbol{R}^3 : \varepsilon^{-1} x Q_n^t - z \in G \subset \boldsymbol{T}^3\},$$

where Q_n is the matrix of rotation in (4.9). For different $z \in Z^3$, $\boldsymbol{T}_{\varepsilon,n}(z)$ are disjoint cubes in $\boldsymbol{R}^3$ with the edge length ε, which are obtained from the "standard" cube $\boldsymbol{T}^3 = [-\frac{1}{2}, \frac{1}{2})^3$ by the mapping

$$\boldsymbol{T}^3 \ni y \quad \longrightarrow \quad x = \varepsilon(y + z)Q_n,$$

and $G_{\varepsilon,n}(z)$ is the image of $G \subset \boldsymbol{T}^3$ by the same mapping. Obviously,

$$|G_{\varepsilon,n}(z)| \le \theta |\boldsymbol{T}_{\varepsilon,n}(z)| = \frac{1}{2} |\boldsymbol{T}_{\varepsilon,n}(z)|. \tag{4.14}$$

Finally, we define

$$G_{k+1} = \bigcup_n \bigcup_{z \in Z_n} G_{\varepsilon,n}(z) \quad \text{where} \quad Z_n = \{z \in Z^3 : \boldsymbol{T}_{\varepsilon,n}(z) \subset G_{k,n}\}. \tag{4.15}$$

Since the components $G_{\varepsilon,n}(z)$ of this set are disjoint, from (4.14) and (4.2) it follows

$$|G_{k+1}| = \sum_n \sum_{z\in Z_n} |G_{\varepsilon,n}(z)| \le \frac{1}{2}\sum_n \sum_{z\in Z_n} |\boldsymbol{T}_{\varepsilon,n}(z)| \le \frac{1}{2}\sum_n |G_{k,n}| = \frac{1}{2}|G_k| \le 2^{-(k+1)}|B_1|;$$

i.e., the estimate (4.2) holds for $k+1$.

In order to get the representation (4.3) for the matrix functions $a^{m,k+1}$, we will use the fact that $x \in G_{\varepsilon,n}(z)$ if and only if $y = \varepsilon^{-1}xQ_n^t - z \in G \subset \boldsymbol{T}^3$. By the periodicity, $\tilde{a}^m(y) \equiv \tilde{a}^m(y+z)$, and hence

$$a^m(\varepsilon^{-1}x) = Q_n^t \tilde{a}^m(\varepsilon^{-1}xQ_n^t)Q_n = Q_n^t \tilde{a}^m(y)Q_n$$

for all $x \in G_{\varepsilon,n}(z)$. Since $a^{0,k} = I$ on $G_k \supset G_{\varepsilon,n}(z)$, we have

$$a^{m,k+1}(x) = (1-\eta_k(x))I + \eta_k(x)\, Q_n^t \tilde{a}^m(y)Q_n \quad \text{on} \quad G_{\varepsilon,n}(z).$$

The properties (3.11) in Lemma 3.5 for $\tilde{a}^m$ imply $\tilde{a}^0 = I$, $\tilde{a}^1 = I + \lambda l^t l$, with $\lambda = \lambda(y)$, $l = l(y)$ satisfying (3.12). Hence

$$a^{0,k+1} = I, \quad a^{1,k+1} = I + \lambda_{k+1}(l^{k+1})^t l^{k+1} \quad \text{on} \quad G_{\varepsilon,n}(z),$$

where

$$\lambda_{k+1}(x) = \eta_k(x)\lambda(y), \quad l^{k+1}(x) = l(y)Q_n, \quad y = \varepsilon^{-1}xQ_n^t - z.$$

Thus λ_{k+1}, l^{k+1} are defined and smooth on $G_{k+1} = \bigcup G_{\varepsilon,n}(z)$. From $\lambda \in C^\infty(\boldsymbol{T}^3)$, $\lambda \equiv 0$ on $\overline{B}_1 \setminus G_{k+1}$, we obtain $\lambda_{k+1} \in C^\infty(\overline{B}_1)$.

The properties (4.4) for λ_{k+1}, l^{k+1} are obvious. In (4.3), we only need to show that

$$a^{0,k+1} = a^{1,k+1} \quad \text{on} \quad B_1 \setminus G_{k+1}. \tag{4.16}$$

This equality on $\boldsymbol{T}_{\varepsilon,n}(z)\setminus G_{\varepsilon,n}(z)$ for $z \in Z_n$ follows from $\tilde{a}^0(y) = \tilde{a}^1(y)$ on $\boldsymbol{T}^3 \setminus G_k$. The remaining points in $B_1 \setminus G_{k+1}$ lie either in $B_1 \setminus G_k$ or in the 2ε-neighborhood of $\partial G_{k,n}$ for some n. Since $\eta_k = 0$ on $B_1 \setminus G_k$ and near $\partial G_{k,n}$, we have (4.16), provided $\varepsilon > 0$ is small enough.

Step 4. We will show that the solutions $u^{m,k+1}$ to the problems (1.6) can be made arbitrarily close to $u^{m,k}$; i.e., for arbitrary constant $h_k > 0$, by the appropriate choice of constants $\delta_k, \mu_k, \varepsilon$ in (4.7), (4.8), (4.12), we have

$$|u^{m,k+1} - u^{m,k}|_0 \le h_k \tag{4.17}$$

for $m = 0$ and 1. For fixed m, k, ε, we introduce the function

$$w = u^{m,k+1} - u^{m,k} + \varepsilon^2\eta_k V, \quad \text{where} \quad V = V(x) = \sum_{i,j} v_{ij}(\varepsilon^{-1}x)D_{ij}u^{m,k}(x),$$

$(v_{ij}) = v^m$ is the matrix function from (4.13), which is smooth on each set $G_{k,n}$. Since $\eta_k = 0$ on $B_1 \setminus G_k$ and near $\partial G_{k,n}$ for each n, the function w is well defined, belongs to $C^\infty(\overline{B_1})$, and $w = 0$ on ∂B_1. Obviously,

$$|u^{m,k+1} - u^{m,k}|_0 \le |w|_0 + \varepsilon^2|V|_0 \le |w|_0 + N\varepsilon^2. \tag{4.18}$$

Here and in the rest of the proof, N denotes different constants not depending on ε and the constants δ_k, μ_k in (4.7), (4.8). By Theorem 2.1,

$$|w|_0 \le N\|F\|_{\mathcal{L}^3(B_1)}, \quad \text{where} \quad F = L^{m,k+1}w. \tag{4.19}$$

Therefore, we will get the estimate (4.17) if we show that the norm of F in $\mathcal{L}^3(B_1)$ can be made arbitrarily small.

On the set $B_1 \setminus G_k$, we have $\eta_k = 0$, $a^{m,k+1} = a^{m,k} = a^{0,k}$; hence

$$F = L^{m,k+1}(u^{m,k+1} - u^{m,k}) = L^{m,k+1}u^{m,k+1} - L^{m,k}u^{m,k} = 0.$$

The estimate of F on G_k is more delicate. We write $F = L^{m,k+1}w = F_1 + F_2$, where

$$F_1(x) = -L^{m,k+1}u^{m,k}(x) + \varepsilon^2\eta_k(x)\sum_{i,j} L^{m,k+1}v^m_{i,j}(\varepsilon^{-1}x)\cdot D_{ij}u^{m,k}(x),$$

$$F_2 = \varepsilon^2\sum_{i,j}\left[L^{m,k+1}(\eta_k v^m_{i,j}(\varepsilon^{-1}x)D_{ij}u^{m,k}(x)) - L^{m,k+1}v^m_{i,j}(\varepsilon^{-1}x)\cdot\eta_k D_{ij}u^{m,k}(x)\right].$$

Using the identity (4.13), we have

$$F_1 = -\sum_{i,j} A^m_{ij}D_{ij}u^{m,k}u = \sum_{i,j}(a^m_{ij} - A^m_{ij})D_{ij}u^{m,k}u \quad \text{on} \quad G'_k = \{\eta_k = 1\} \subset G_k.$$

By virtue of (4.7), (4.8),

$$|F_1| \le N\delta_k \quad \text{on } G'_k, \qquad |F_1| \le N \quad \text{on } G_k \setminus G'_k.$$

Hence the norm of F_1 in $\mathcal{L}^3(B_1)$ is bounded by $N(\delta_k + \mu^{\frac{1}{3}})$ and can be made arbitrarily small. The same is true for $F = F_1 + F_2$ because of the estimate $|F_2| \le N_1\varepsilon$ with a constant N_1, which does not depend on ε. Now the desired approximation (4.17) follows from (4.19).

Step 5. Using (4.5) and (4.17), it is easy to complete our construction. We choose a sequence $\{h_k\}$ satisfying $\sum h_k \le \frac{1}{4}H$. By the Cauchy criterion, sequences $\{u^{0,k}\}, \{u^{1,k}\}$ converge in $C(\overline{B_1})$ to some functions u^0, u^1:

$$|u^{0,k} - u^0|_0,\ |u^{1,k} - u^1|_0 \ \to 0 \quad \text{as} \quad k \to \infty.$$

By the triangle inequality, for any natural K,

$$|u^{0,K} - u^{1,K}|_0 \ge |u^{0,0} - u^{1,0}|_0 - \sum_{k=0}^{K-1}|u^{0,k+1} - u^{0,k}|_0 - \sum_{k=0}^{K-1}|u^{1,k+1} - u^{1,k}|_0$$

$$\ge H - 2\sum_{k=0}^{\infty} h_k \ge \frac{1}{2}H > 0.$$

Therefore,

$$|u^0 - u^1|_0 = \lim_{K\to\infty}|u^{0,K} - u^{1,K}|_0 \ge \frac{1}{2}H > 0.$$

Theorem 1.2 is completely proved. □

Acknowledgments. The author is grateful to N. S. Nadirashvili for an opportunity to learn about his result before its publication, to N. V. Krylov and P. Manselli for very useful discussions on the topics of this paper, and to Yu Yuan for corrections of misprints. The author is also deeply indebted to the late Eugene Fabes for his valuable suggestions and critical remarks.

REFERENCES

[1] A. D. Aleksandrov, *Uniqueness conditions and estimates for the solutions of the Dirichlet problem*, Vestnik Leningrad Univ., 18 (1963), pp. 5–29 (in Russian); Amer. Math. Soc. Trans., 68 (1968), pp. 89–119 (in English).

[2] A. D. Aleksandrov, *The impossibility of general estimates for solutions and of uniqueness conditions for linear equations with norms weaker than in L_n*, Vestnik Leningrad Univ., 18 (1966), pp. 5–29 (in Russian); Amer. Math. Soc. Trans., 68 (1968), pp. 89–119 (in English).

[3] O. Arena and P. Mancelli, *A class of elliptic operators in $\boldsymbol{R}^3$ in non-divergent form with measurable coefficients*, Le Matematiche, 48 (1993), pp. 163–177.

[4] R. F. Bass, *The Dirichlet problem for radially homogeneous elliptic operators*, Trans. Amer. Math. Soc., 320 (1990), pp. 593–614.

[5] R. F. Bass and E. Pardoux, *Uniqueness for diffusions with piecewise constant coefficients*, Probab. Theory Related Fields, 76 (1987), pp. 557–572.

[6] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland, Amsterdam, New York, Oxford, 1978.

[7] L. Bers, F. John, and M. Schecter, *Partial Differential Equations*, Interscience, New York, London, Sydney, 1964.

[8] 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.

[9] M. C. Cerutti, L. Escauriaza, and E. B. Fabes, *Uniqueness for some diffusions with discontinuous coefficients*, Ann. Probab., 19 (1991), pp. 525–537.

[10] M. C. Cerutti, L. Escauriaza, and E. B. Fabes, *Uniqueness in the Dirichlet problem for some elliptic operators with discontinuous coefficients*, Ann. Mat. Pura Appl. (4) Ser. A, 163 (1993), pp. 161–180.

[11] M. C. Cerutti, E. B. Fabes, and P. Manselli, *Uniqueness for elliptic operators with time-independent coefficients*, Pitman Res. Notes Math. Ser., 350, Longman, Harlow, 1996, pp. 112–135.

[12] H. O. Cordes, *Über die erste Randwertaufgabe bei quasilinearen Differentialgleichungen zweiter Ordnung in mehr als zwei Variablen*, Math. Ann., 131 (1956), pp. 278–312.

[13] M. I. Freidlin, *Dirichlet's problem for an equation with periodic coefficients depending on a small parameter*, Teor. Veroyatnost i Primenen., 9 (1964), pp. 133–138 (in Russian); Theory Probab. Appl., 9 (1964), pp. 121–125 (in English).

[14] D. Gilbarg and J. Serrin, *On isolated singularities of solutions of second order elliptic differential equations*, J. Anal. Math., 4 (1955/56), pp. 309–340.

[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[16] R. R. Jensen, *Uniformly elliptic PDEs with bounded, measurable coefficients*, J. Fourier Anal. Appl., 2 (1996), pp. 237–259.

[17] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer-Verlag, Berlin, New York, 1994.

[18] N. V. Krylov, *Nonlinear Elliptic and Parabolic Equations of Second Order*, Nauka, Moscow, 1985 (in Russian); Reidel, Dordrecht, 1987 (in English).

[19] N. V. Krylov, *On Itô's stochastic integral equations*, Teor. Veroyatnost i Primenen., 14 (1969), pp. 340–348 (in Russian); Theory Probab. Appl., 14 (1969), pp. 330–336 (in English).

[20] N. V. Krylov, *On the selection of a Markov process from a system of processes and the construction of quasi-diffusion processes*, Izv. Akad. Nauk SSSR, ser. Matem., 37 (1973), pp. 691–708 (in Russian); Math. USSR Izvestija, 7 (1973), pp. 691–709 (in English).

[21] N. V. Krylov, *Once more about the connection between elliptic operators and Itô's stochastic equations*, in Statistics and Control of Stochastic Processes, Moscow, 1984, pp. 214–229 (in Russian); Optimization Software, New York, 1985 (in English).

[22] N. V. Krylov, *On one-point uniqueness for elliptic equations*, Comm. Partial Differential Equations, 17 (1992), pp. 1759–1784.

[23] N. V. Krylov and M. V. Safonov, *A certain property of solutions of parabolic equations with measurable coefficients*, Izv. Akad. Nauk SSSR, Ser. Mat., 44 (1980), pp. 161–175 (in Russian); Math. USSR Izvestija, 16 (1981), pp. 151–164 (in English).

[24] O. A. Ladyzhenskaya and N. N. Ural'tseva, *Linear and Quasilinear Elliptic Equations*, Nauka, Moscow, 1964 (in Russian); Academic Press, New York, 1968 (in English); 2nd Russian ed. 1973.

[25] N. S. Nadirashvili, *Nonuniqueness in the martingale problem and the Dirichlet problem for uniformly elliptic operators*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 24 (1997), pp. 537–550.

[26] C. Pucci, *Limitazioni per soluzioni di equazioni ellittiche*, Ann. Mat. Pura Appl., 74 (1966), pp. 15-30.

[27] M. V. Safonov, *Harnack inequality for elliptic equations and the Hölder property of their solutions*, Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov (LOMI), 96 (1980), pp. 272–287 (in Russian); J. Soviet Math., 21 (1983), pp. 851–863 (in English).

[28] M. V. Safonov, *Unimprovability of estimates of Hölder constants for solutions of linear elliptic equations with measurable coefficients*, Mat. Sb., 132 (1987), pp. 272–288 (in Russian); Math. USSR Sbornik, 60 (1988), pp. 269–281 (in English).

[29] M. V. Safonov, *On a weak uniqueness for some elliptic equations*, Comm. Partial Differential Equations, 19 (1994), pp. 943–957.

[30] D. W. Stroock and S. R. S. Varadhan, *Multidimensional Diffusion Processes*, Springer-Verlag, Berlin, Heidelberg, New York, 1979.

[31] D. W. Stroock and S. R. S. Varadhan, *Diffusion processes with continuous coefficients*, I *and* II, Comm. Pure Appl. Math., 22 (1969), pp. 345–400, 479–530.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 896–911

A DENSENESS THEOREM WITH AN APPLICATION TO A TWO-DIMENSIONAL INVERSE POTENTIAL REFRACTION PROBLEM*

BERND HOFMANN†

Abstract. In this paper we will be concerned with the problem of reconstructing a region $D \subset \Omega \subset \mathbb{R}^2$, from the knowledge of the boundary potential $u|_{\partial\Omega}$, where u satisfies

$$\operatorname{div}((\chi_{\bar{\Omega}\setminus D} + a\chi_D)\nabla u) = 0 \quad \text{in } \Omega \qquad \text{and} \qquad \frac{\partial u}{\partial \nu} = I \quad \text{on } \partial\Omega,$$

with a a real, positive constant. We will show that the domain derivative of the corresponding forward mapping is injective. This is done by proving the denseness of a certain subspace of $L^1(\partial D)$. The novelty of the result is that our proof is valid without any restriction on the Neumann boundary data I.

Key words. parameter identification, potential refraction, Hilbert problems

AMS subject classifications. 31A25, 35R30

PII. S0036141098336108

1. Introduction.

The denseness theorem proved in this paper is related to an inverse problem in electrical impedance tomography. Its relation to this problem was first worked out in [2], [3]. Here we will briefly outline its relevance without any proofs.

We start by introducing some notations. If $D \subset \mathbb{R}^n$ is bounded with smooth boundary ∂D, we will denote the outward unit normal of ∂D by ν or, to avoid ambiguities, more precisely by ν_D. The Euclidean scalar product in $\mathbb{R}^n$ is denoted by $\langle \cdot, \cdot \rangle$. The symbols for the standard function spaces will always stand for the spaces of real valued functions. A lower index $\mathbb{C}$ will be added for the corresponding space of complex valued functions. In particular, $C(D)$ and $C(D)_{\mathbb{C}}$ denote the spaces of real and complex valued, continuous functions on D, respectively. For a nonnegative integer k and $0 < \alpha < 1$, we set

$$\mathcal{P}^{k,\alpha}(D) := \{f \in C^{k,\alpha}(\bar{D}) : f \text{ is harmonic in } D\},$$
$$\mathfrak{X}^{k,\alpha}(\partial D) := C^{k,\alpha}(\partial D, \mathbb{R}^n).$$

The spaces $\mathfrak{X}^{k,\alpha}(\partial D)$ will be interpreted as vector fields on ∂D. If $Z \in \mathfrak{X}^{k,\alpha}(\partial D)$, then we will denote its normal component $\langle Z, \nu \rangle$ by Z_ν. As usual, the upper indices k and α will be dropped if they are zero. In the two-dimensional case $D \subset \mathbb{R}^2$ we will identify $\mathbb{R}^2 \simeq \mathbb{C}$ via $(x, y) \simeq x + iy$ and set

$$\mathcal{H}^{\alpha}(D) := \{f \in C^{\alpha}(\bar{D})_{\mathbb{C}} : f \text{ is holomorphic in } D\}.$$

Further, the unit tangent on ∂D with positive orientation will be denoted by τ or τ_D.

*Received by the editors March 20, 1998; accepted for publication July 23, 1998; published electronically June 3, 1999. This research was supported by the Deutsche Forschungsgemeinschaft.
http://www.siam.org/journals/sima/30-4/33610.html

†Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestraße 16–18, D-37083 Göttingen, Germany (bhofmann@math.uni-goettingen.de).

Let $\Omega \subset \mathbb{R}^n$ be an open bounded set with smooth boundary $\partial\Omega$. Consider the elliptic boundary value problem

$$\operatorname{div}(\sigma \nabla u) = 0 \qquad \text{in } \Omega, \tag{1.1a}$$

$$\sigma \frac{\partial u}{\partial \nu} = I \qquad \text{on } \partial\Omega, \tag{1.1b}$$

$$\int_{\partial\Omega} u \, ds = 0. \tag{1.1c}$$

Here $\partial/\partial\nu$ denotes the outward normal derivative on $\partial\Omega$, div the divergence, and ∇ the gradient. In the applications σ is a positive electrical conductivity, I a boundary current satisfying $\int_{\partial\Omega} I \, ds = 0$, and u the potential of the electric field generated by I. A suitable weak formulation of (1.1) can be shown to have a unique solution.

We will exclusively consider the case in which $I \in C(\partial\Omega)$ and the conductivity is of the form

$$\sigma = \chi_{\bar{\Omega}\setminus D} + a\chi_D,$$

where a is a real positive constant, D is an open subset with C^2-smooth boundary such that $\bar{D} \subset \Omega$, and χ_M denotes the characteristic function of a subset $M \subset \mathbb{R}^n$. Then the solution u of (1.1) is in $C(\bar{\Omega}) \cap C^2(\Omega \setminus \partial D)$. Setting $u_+ := u|_{\Omega\setminus D}$ and $u_- := u|_D$, (1.1) can be rewritten in the form

$$\triangle u = 0 \qquad \text{in } \Omega \setminus \partial D, \tag{1.2a}$$

$$\frac{\partial u_+}{\partial \nu} - a\frac{\partial u_-}{\partial \nu} = 0 \qquad \text{on } \partial D, \tag{1.2b}$$

$$\frac{\partial u_+}{\partial \nu} = I \qquad \text{on } \partial\Omega, \tag{1.2c}$$

$$\int_{\partial\Omega} u_+ \, ds = 0. \tag{1.2d}$$

For the remainder of this section we will assume that ∂D is $C^{2,\alpha}$-smooth for some $0 < \alpha < 1$. Then the first and second derivatives of u_- and u_+ can be extended Hölder-continuously onto ∂D.

We are interested in the inverse problem of reconstructing the conductivity inside Ω from the boundary potential $u|_{\partial\Omega}$. It can be seen from simple counterexamples that, in general, it is not possible to reconstruct D and a simultaneously from a single pair $I, u|_{\partial\Omega}$. Therefore it is assumed that the conductivity constant a is known. Then the conductivity depends only on the region D, and we can define the forward mapping

$$F : D \mapsto u|_{\partial\Omega}.$$

This mapping can be linearized at D in the following manner. For Z in some sufficiently small neighborhood U of 0 in $\mathfrak{X}^{2,\alpha}(\partial D)$ the set

$$\{x + Z(x) : x \in \partial D\}$$

is again the $C^{2,\alpha}$-smooth boundary of an open set D_Z. For a compact subset $K \subset \bar{\Omega} \setminus \partial D$ we can choose U small enough such that $K \cap \bar{D}_Z = \emptyset$ for any $Z \in U$. We then may consider the mapping

$$\Psi_K : \mathfrak{X}^{2,\alpha}(\partial D) \supset U \to C(K), \quad Z \mapsto u_Z|_K, \tag{1.3}$$

where u_Z is the solution of (1.2) with conductivity $\chi_{\bar\Omega\setminus D_Z}+a\chi_{D_Z}$, so that $\Psi_{\partial\Omega}(D_Z)=F(D_Z)$.

THEOREM 1.1. *For each compact $K\subset\bar\Omega\setminus\partial D$ the mapping* (1.3) *is Fréchet differentiable. Exhausting $\bar\Omega\setminus\partial D$ by compact sets, for $Z\in\mathfrak{X}^{2,\alpha}(\partial D)$ we can define the differentiated potential*

$$u' := \frac{d}{ds}\bigg|_{s=0}\Psi_K(D_{sZ})$$

as a function in $C(\bar\Omega\setminus\partial D)$. Then u'_+ and u'_- together with their first partial derivatives can be continuously extended to ∂D. If u is the solution of (1.2) *with respect to the region D, then u' is the unique solution of the transmission problem*

$$\triangle u' = 0 \qquad \text{in } \Omega\setminus\partial D, \tag{1.4a}$$

$$u'_+ - u'_- = -Z_\nu\Big(\frac{\partial u_+}{\partial\nu}-\frac{\partial u_-}{\partial\nu}\Big) \qquad \text{on } \partial D, \tag{1.4b}$$

$$\frac{\partial u'_+}{\partial\nu} - a\frac{\partial u'_-}{\partial\nu} = (1-a)\,\mathrm{Div}(Z_\nu\,\mathrm{Grad}\,u) \qquad \text{on } \partial D, \tag{1.4c}$$

$$\frac{\partial u'_+}{\partial\nu} = 0 \qquad \text{on } \partial\Omega, \tag{1.4d}$$

$$\int_{\partial\Omega} u'_+\,ds = 0. \tag{1.4e}$$

Here Div *and* Grad *denote the surface divergence and surface gradient on ∂D, which, in the two-dimensional case, may both be replaced by the derivative with respect to the arc length.*

Theorem 1.1 can be obtained by representing the solution of (1.2) by single layer potentials over the boundaries $\partial\Omega$ and ∂D and then using results by Potthast [7], [8] on the Fréchet differentiability of the classical boundary integral operators. A proof using weak formulations for the analogous result, when Dirichlet data $u|_{\partial\Omega}$ are prescribed and Neumann data $\partial u/\partial\nu$ on $\partial\Omega$ are measured, can be found in [4].

Denoting by $\mathfrak{X}^{2,\alpha}_T(\partial D)$ the space of vector fields in $\mathfrak{X}^{2,\alpha}(\partial D)$, which are tangential on ∂D, the above Fréchet derivative of F at D may be viewed as a mapping

$$F'_D : \mathfrak{X}^{2,\alpha}(\partial D)/\mathfrak{X}^{2,\alpha}_T(\partial D) \to C(\partial\Omega), \quad [Z]\mapsto u'|_{\partial\Omega}. \tag{1.5}$$

The uniqueness of the linearized inverse problem amounts to proving the injectivity of F'_D. Again, it can be seen from counterexamples that a general uniqueness result is false, unless $\mathbb{R}^n\setminus D$ is assumed to be connected. If we suppose that this condition is fulfilled, the injectivity of F'_D can then be obtained from the denseness of a certain subspace of $C^\alpha(\partial D)$ in $L^1(\partial D)$. This is one of the main results of [2], and it is proved without using the characterization of the derivative in Theorem 1.1.

THEOREM 1.2. *Assume that $\mathbb{R}^n\setminus D$ is connected and let the notations be as in Theorem* 1.1. *If Z is in the kernel of F'_D and $a\neq 1$, then for any $v\in\mathcal{P}^{1,\alpha}(D)$ we have*

$$\int_{\partial D} Z_\nu\langle\nabla u_+,\nabla v\rangle\,ds = 0. \tag{1.6}$$

Proof. The trivial Cauchy data $F'_D(Z)=u'_+|_{\partial\Omega}=0$ and (1.4d) imply $u'_+=0$ by Holmgren's uniqueness theorem. Now the statement follows by plugging the Cauchy

data of u'_- from (1.4b), (1.4c) into Green's second theorem applied to u'_- and v in D. □

For $a > 0$ define a refraction operator on the space of vector fields on ∂D by

$$R_a^{\partial D} : \mathfrak{X}(\partial D) \to \mathfrak{X}(\partial D), \quad X \mapsto X - (a-1)\langle X, \nu\rangle \nu. \tag{1.7}$$

The operator $R_a^{\partial D}$ multiplies the normal component of X by a and leaves the tangential component unchanged. Clearly, the zeros of X and $R_a^{\partial D}(X)$ coincide. Set $\nabla_a f = R_a^{\partial D}(\nabla f|_{\partial D})$, $f \in C^1(\bar{D})$. The main result can now be formulated as follows.

THEOREM 1.3. *Let $D \subset \mathbb{R}^2$ be an open, bounded, connected, and simply connected set with C^2-smooth boundary. If $u \in \mathcal{P}^{1,\alpha}(D)$ is not constant and $a > 0$, then the function space*

$$\mathcal{F}_{\partial D}(\nabla_a u) := \{\langle \nabla_a u, \nabla v\rangle : v \in \mathcal{P}^{1,\alpha}(D)\} \subset C^\alpha(\partial D) \tag{1.8}$$

is dense in $L^1(\partial D)$.

Observing that for the solution of (1.2) we have from (1.2b) and the continuity of u

$$\nabla u_+|_{\partial D} = R_a^{\partial D}(\nabla u_-|_{\partial D}) = \nabla_a u_-|_{\partial D},$$

the injectivity of F'_D in the two-dimensional case is now almost immediate.

COROLLARY 1.4. *Let D be an open, bounded, connected, and simply connected set with $C^{2,\alpha}$-smooth boundary and assume that $I \neq 0$ and $a > 0$, $a \neq 1$. Then the Fréchet derivative* (1.5) *is injective.*

The denseness of the spaces $\mathcal{F}(\nabla_a u)$ is well known if the vector field $\nabla_a u$ has no zeros. This can be deduced from the standard theory of Hilbert problems, as will be explained in the next section. The results in [9], [1] on the local uniqueness of F are restricted to injections I which ensure $\nabla u_+(x) \neq 0$, $x \in \partial D$ independently of D. In Theorem 1.3 we have not excluded vector fields with zeros on ∂D and therefore we need no restrictions on I in Corollary 1.4.

2. Sketch of the proof. Obviously, $h \in C^\alpha(\partial D)$ is an element of $\mathcal{F}_{\partial D}(\nabla_a u)$, if the oblique derivative problem

$$\text{Find } v \in \mathcal{P}^{1,\alpha}(D) \text{ satisfying } \langle \nabla_a u, \nabla v\rangle = h \text{ on } \partial D \tag{2.1}$$

is solvable. In arbitrary dimensions a solution theory for the Laplace equation with boundary condition $\langle X, \nabla v\rangle = h$ on ∂D, where X is some vector field on ∂D, is available only if the normal component of X has no zeros on ∂D (see [11]). Since u is harmonic in D, we have

$$\int_{\partial D} \langle \nabla_a u, \nu\rangle \, ds = a \int_{\partial D} \partial u/\partial \nu \, ds = 0,$$

and therefore this condition will never be satisfied for (2.1). In two dimensions oblique derivative problems are more thoroughly investigated through their equivalent formulation as Hilbert problems for holomorphic functions (see [6]). This provides a solution theory for the case where X has no zeros on ∂D. Unfortunately, this condition in general also will not be satisfied for (2.1). However, unlike the normal derivative, the gradient of a nonconstant harmonic function can vanish only on a set of measure zero on ∂D. Using this and the multiplicative structure on $\mathbb{R}^2 \simeq \mathbb{C}$, it is possible to

show the solvability of (2.1) for *enough inhomogeneities* h by considering a suitable modification of (2.1). On the other hand, this procedure restricts the proof to the two-dimensional case.

For $z \in \mathbb{C}$ we denote the real and imaginary part by $\operatorname{Re} z$ and $\operatorname{Im} z$, respectively. The complex conjugate of z is denoted by $\overline{z}$.

If f is holomorphic in D, then $\operatorname{Re} f$ and $\operatorname{Im} f$ are harmonic in D. If u is harmonic in D, then ∇u is antiholomorphic, i.e., $\partial u/\partial x_1 - i\partial u/\partial x_2$ is holomorphic in D. Since D is assumed to be simply connected, the converse is also true. If f is holomorphic in D, then there exists a harmonic function u on D with $\overline{f} \simeq \nabla u$. This sets up a one-to-one correspondence between $\mathcal{H}^\alpha(D)$ and the gradients of functions in $\mathcal{P}^{1,\alpha}(D)$. If $z, w \in \mathbb{C}$, then $\operatorname{Re}(z\overline{w})$ is equal to the Euclidean scalar product of z and w, both interpreted as vectors in $\mathbb{R}^2$.

Now let $u, v \in \mathcal{P}^{1,\alpha}(D)$ and define $f, g \in \mathcal{H}^\alpha(D)$ by $\nabla u \simeq \overline{f}$ and $\nabla v \simeq \overline{g}$. Define the refraction operator (1.7) on the complex functions on ∂D through the identification $\mathbb{R}^2 \simeq \mathbb{C}$, and set

$$\varphi_a := \overline{R_a^{\partial D}(\overline{f}|_{\partial D})} \in C^\alpha(\partial D)_{\mathbb{C}}. \tag{2.2}$$

Then we have $\nabla_a u \simeq \overline{\varphi}_a$. From the above we find $\operatorname{Re}(\overline{\varphi}_a g) = \langle \nabla_a u, \nabla v \rangle$, hence

$$\mathcal{F}_{\partial D}(\nabla_a u) = \mathcal{F}^c_{\partial D}(\varphi_a), \tag{2.3}$$

where

$$\mathcal{F}^c_{\partial D}(\varphi_a) := \{\operatorname{Re}(\overline{\varphi}_a g) : g \in \mathcal{H}^\alpha(D)\}. \tag{2.4}$$

Now we can reformulate Theorem 1.3 in terms of holomorphic functions.

THEOREM 2.1. *Assume that $f \in \mathcal{H}^\alpha(D)$ is not identically zero and that $a > 0$, and define φ_a as in* (2.2). *Then the function space $\mathcal{F}^c_{\partial D}(\varphi_a)$ is dense in $L^1(\partial D)$.*

The oblique derivative problem (2.1) is equivalent to the Hilbert problem

$$\text{Find } g \in \mathcal{H}^\alpha(D) \text{ satisfying } \operatorname{Re}(\overline{\varphi}_a g) = h \text{ on } \partial D. \tag{2.5}$$

Therefore Theorems 1.3 and 2.1 state that the inhomogeneities $h \in C^\alpha(\partial D)$ admitting a solution of (2.5) are dense in $L^1(\partial D)$. This denseness is fairly simple to show, if $\varphi_a(z) \neq 0$ for all $z \in \partial D$. For this we need the notion of the index of φ_a defined as the integer

$$\operatorname{Ind}(\varphi_a) = \frac{1}{2\pi} \arg \varphi_a \Big|_{\partial D}.$$

The Hilbert problem is solvable for all $h \in C^\alpha(\partial D)$ if and only if the index is non-negative (see [6]). The functions

$$\varphi_\lambda = \overline{R_\lambda^{\partial D}(\overline{f}|_{\partial D})}, \qquad \lambda = (1-t)a + t, \quad t \in [0,1],$$

have no zeros for all t. Therefore a continuity argument implies

$$\operatorname{Ind}(\varphi_a) = \operatorname{Ind}(f|_{\partial D}). \tag{2.6}$$

Since f is holomorphic, the index of $f|_{\partial D}$ is equal to the number of zeros of f in D. We thus have $\operatorname{Ind}(\varphi_a) \geqslant 0$, and therefore $C^\alpha(\partial D) \subset \mathcal{F}^c_{\partial D}(\varphi_a)$, which clearly implies the asserted $L^1(\partial D)$-denseness.

In [3] it is shown that the relation between the index of the boundary values of a holomorphic function and its zeros inside the region is still valid for the sectionally holomorphic function f_u with $\overline{f}_u|_D = \nabla u_-$, $\overline{f}_u|_{\Omega\setminus\bar{D}} = \nabla u_+$, provided that ∂D is analytic. In [9] Powell has proved under suitable restrictions on the Neumann data I that$\nabla u_+|_{\partial\Omega}$ has no zeros and its index is zero. By the above argument, this implies $\nabla u_+|_{\partial D} \neq 0$, whence the denseness of $\mathcal{F}_{\partial D}(\nabla_a u_-)$ follows. In [9] the condition $\nabla u_+(z) \neq 0$, $z \in \partial D$, is used to derive local injectivity of F. These results are generalized in [1].

For general currents I, there is no guarantee that ∇u_+ has no zeros on ∂D. If we have $\varphi_a(z) = 0$ for some $z \in \partial D$, then it is obvious that (2.5) can be solvable only if the inhomogeneity h also vanishes in z. Thus in this case $C^\alpha(\partial D) \not\subset \mathcal{F}^c_{\partial D}(\varphi_a)$, and therefore the difficulty in the proof of Theorem 2.1 is to work around the zeros of φ_a.

The proof, presented in the next section, consists of the following steps:

(i) With the aid of the Riemann mapping theorem, we will show that without loss of generality we may assume that

$$D = \mathcal{D} := \{z \in \mathbb{C} : |z| < 1\}, \qquad \partial D = \mathbf{S}^1 := \{z \in \mathbb{C} : |z| = 1\}.$$

(ii) We divide the boundary condition in (2.5) by $|\varphi_1|^2$, $\varphi_1 := \overline{R_1^{\mathbf{S}^1}(\overline{f}|_{\mathbf{S}^1})} = f|_{\mathbf{S}^1}$ and obtain

$$\operatorname{Re}(\overline{\varphi}_a g)/|\varphi_1|^2 = \operatorname{Re}\left(\overline{\varphi_a/\varphi_1}\cdot g/\varphi_1\right) = h/|\varphi_1|^2. \tag{2.7}$$

Noting that $\mathcal{F}^c_{\partial D}(\varphi_a)$ is a linear subspace of $L^1(\mathbf{S}^1)$, it is clear that the solvability of (2.5) need be shown only for some subset $\mathcal{I} \subset C^\alpha(\mathbf{S}^1)$ whose linear span is a dense subspace of $L^1(\mathbf{S}^1)$. We choose this subset so that the supports of its functions do not intersect with a neighborhood of the zeros of φ_1 which, by definition, coincide with those of φ_a. Then the right-hand side in (2.7) can be considered a function in $C^\alpha(\mathbf{S}^1)$ by setting it equal to zero outside the support of h. For a function φ on $\mathbf{S}^1$ we will denote the set of its zeros by $\mathrm{N}(\varphi)$ and define a set of suitable inhomogeneities for (2.5) by setting

$$C^\alpha(\mathbf{S}^1, J) := \{h \in C^\alpha(\mathbf{S}^1) : \operatorname{supp} h \subset J\}, \qquad J \subset \mathbf{S}^1, \tag{2.8}$$

$$\mathcal{I}(\varphi_a) := \bigcup_{\substack{J\subset\mathbf{S}^1\setminus\mathrm{N}(\varphi_a)\\ J \text{ closed, connected}}} C^\alpha(\mathbf{S}^1, J). \tag{2.9}$$

The span of $\mathcal{I}(\varphi_a)$ is dense in $L^1(\mathbf{S}^1)$.

If f is assumed to admit an extension to a holomorphic function in some neighborhood of $\bar{\mathcal{D}}$, then the set $\mathrm{N}(\varphi_1) = \mathrm{N}(f|_{\mathbf{S}^1})$ is finite. We prove that in this case $\psi := \varphi_a/\varphi_1$ can be extended into $\mathrm{N}(\varphi_1)$ as an analytic function ψ on $\mathbf{S}^1$. Furthermore, ψ vanishes nowhere on $\mathbf{S}^1$ and the index of ψ is zero. Thus the Hilbert problem

$$\text{Find } g \in \mathcal{H}^\alpha(\mathcal{D}) \text{ satisfying } \operatorname{Re}(\overline{\psi}g) = h \text{ on } \mathbf{S}^1 \tag{2.10}$$

is solvable for all $h \in C^\alpha(\mathbf{S}^1)$. Now if $h \in \mathcal{I}(\varphi_a)$ and $\tilde{g} \in \mathcal{H}^\alpha(\mathcal{D})$ is a solution of $\operatorname{Re}(\overline{\psi}\tilde{g}) = h/|\varphi_1|^2$ on $\mathbf{S}^1$, then $g := f\tilde{g} \in \mathcal{H}^\alpha(\mathcal{D})$ satisfies

$$\operatorname{Re}(\overline{\varphi}_a g) = \operatorname{Re}(\overline{\psi\varphi}_1 \cdot \varphi_1\tilde{g}) = |\varphi_1|^2 \operatorname{Re}(\overline{\psi}\tilde{g}) = h \text{ on } \mathbf{S}^1. \tag{2.11}$$

Thus if f is holomorphic in a neighborhood of $\bar{\mathcal{D}}$ we have

$$\mathcal{I}(\varphi_a) \subset \mathcal{F}^c_{\mathbf{S}^1}(\varphi_a), \tag{2.12}$$

hence for this case Theorem 2.1 is proven.

Note that in step (ii), we will make no use of $D = \mathcal{D}$. Only the assumption that ∂D is analytic will be needed. In this case it follows from elliptic regularity results that the interior potential u_- in (1.2) can be extended as a harmonic function in a neighborhood of $\bar{D}$. Thus for analytic boundaries Corollary 1.4 already can be deduced at this point of the proof.

The formulation of step (ii) with $D = \mathcal{D}$ is motivated by the fact that there exist explicit formulas for the solution of (2.10) for the unit disk. These formulas will be used in the next step.

(iii) Finally, we treat the general case $f \in \mathcal{H}^\alpha(\mathcal{D})$. For $r < 1$ we set

$$f_r(z) := f(rz), \quad |z| < 1/r, \qquad \varphi_a^{(r)} := \overline{R_a^{\mathbf{S}^1}(\overline{f}_r|_{\mathbf{S}^1})}, \quad \varphi_1^{(r)} := f_r|_{\mathbf{S}^1}.$$

Obviously, we have

$$\varphi_a^{(r)} \xrightarrow[r\to 1]{} \varphi_a, \quad \varphi_1^{(r)} \xrightarrow[r\to 1]{} \varphi_1 \qquad (C^\alpha(\mathbf{S}^1)_{\mathbb{C}}\text{-convergence}). \tag{2.13}$$

Since the functions f_r are holomorphic in a neighborhood of $\bar{\mathcal{D}}$, from step (ii) we have

$$\mathcal{I}(\varphi_a^{(r)}) \subset \mathcal{F}^c_{\mathbf{S}^1}(\varphi_a^{(r)}).$$

Let $h \in \mathcal{I}(\varphi_a)$ and $J \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ be closed and connected with $\operatorname{supp} h \subset J$. By definition, φ_1 has no zeros in J. From the uniform convergence $\varphi_1^{(r)} \to \varphi_1$ it follows that there exists $r_0 < 1$, such that $\varphi_1^{(r)}$ has no zeros in J for $r_0 < r < 1$. Therefore, for these r we have $h \in \mathcal{I}(\varphi_a^{(r)})$. Consequently, there exist $g_r \in \mathcal{H}^\alpha(\mathcal{D})$ with boundary values satisfying

$$\mathrm{Re}(\overline{\varphi_a^{(r)}} g_r) = h, \qquad r_0 < r < 1. \tag{2.14}$$

Since the $\varphi_a^{(r)}$ approximate φ_a, we intend to show that

$$\mathrm{Re}(\overline{\varphi}_a g_r) \xrightarrow[r\to 1]{} h \qquad (L^1(\mathbf{S}^1)\text{-convergence}). \tag{2.15}$$

This establishes that $\mathcal{I}(\varphi_a)$ is contained in the closure of $\mathcal{F}^c_{\mathbf{S}^1}(\varphi_a)$ in $L^1(\mathbf{S}^1)$, which completes the proof of Theorem 2.1.

Bearing in mind that $\mathrm{Re}(\overline{z}w)$, $z, w \in \mathbb{C}$, is the Euclidean scalar product of z and w, we obtain the estimate

$$\begin{aligned}
\|\mathrm{Re}(\overline{\varphi}_a g_r) - h\|_{L^1(\mathbf{S}^1)} &= \left\|\mathrm{Re}(\overline{\varphi}_a g_r) - \mathrm{Re}(\overline{\varphi_a^{(r)}} g_r)\right\|_{L^1(\mathbf{S}^1)} \\
&= \left\|\mathrm{Re}\left((\overline{\varphi}_a - \overline{\varphi_a^{(r)}}) g_r\right)\right\|_{L^1(\mathbf{S}^1)} \\
&\leqslant \|\varphi_a - \varphi_a^{(r)}\|_{L^\infty(\mathbf{S}^1)_{\mathbb{C}}} \|g_r\|_{L^1(\mathbf{S}^1)_{\mathbb{C}}}.
\end{aligned}$$

From (2.13) we have $\|\varphi_a - \varphi_a^{(r)}\|_{L^\infty(\mathbf{S}^1)_{\mathbb{C}}} \to 0$. Thus, for the proof of (2.15), we have to show that the functions g_r can be chosen such that the restrictions $g_r|_{\mathbf{S}^1}$ are uniformly $L^1(\mathbf{S}^1)_{\mathbb{C}}$-bounded for $r \to 1$.

3. Proof of Theorem 2.1. Let D be as in Theorem 2.1 and $f \in \mathcal{H}^\alpha(D)$. The Riemann mapping theorem asserts the existence of a $C^{1,\alpha}$-diffeomorphism

$$\Phi : \bar{D} \to \bar{\mathcal{D}}$$

such that $\Phi|_D : D \to \mathcal{D}$ is biholomorphic. Let $u \in \mathcal{P}^{1,\alpha}(D)$ with $\nabla u \simeq \overline{f}$. Then $u^* := u \circ \Phi^{-1} \in C^{1,\alpha}(\bar{\mathcal{D}})$ is harmonic in $\mathcal{D}$. Let $f^* \in \mathcal{H}^\alpha(\mathcal{D})$ be defined by $\nabla u^* \simeq \overline{f}$,

$$\varphi_a := \overline{R_a^{\partial D}(\overline{f}|_{\partial D})}, \quad \text{and} \quad \varphi_a^* := \overline{R_a^{\mathbf{S}^1}(\overline{f^*}|_{\mathbf{S}^1})} \in C^\alpha(\mathbf{S}^1).$$

As usual, we denote the complex derivative of Φ by Φ'. If $z \in \mathbb{C}$, $z = a + ib$, $a, b \in \mathbb{R}$, then the multiplication by z under the identification $\mathbb{R}^2 \simeq \mathbb{C}$ is equivalent to the multiplication by the matrix $\left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$. Therefore we have

$$\det(J\Phi) = |\Phi'|^2,$$

where $J\Phi$ denotes the Jacobi matrix of Φ. In particular, Φ preserves the orientation and Φ' has no zeros. The multiplication of a line vector in $\mathbb{R}^2$ by $\left(\begin{smallmatrix} a & -b \\ b & a \end{smallmatrix}\right)$ from the right corresponds to multiplication by $\overline{z}$. Define

$$\Phi^* : L^1(\mathbf{S}^1) \to L^1(\partial D), \quad \phi \mapsto \phi \circ \Phi,$$

and for $q \in C(\partial D)$ let M_q be the multiplication mapping

$$M_q : L^1(\partial D) \to L^1(\partial D), \quad \psi \mapsto q \cdot \psi.$$

LEMMA 3.1. *The functions φ_a, φ_a^* and the spaces $\mathcal{F}^c_{\partial D}(\varphi_a)$, $\mathcal{F}^c_{\mathbf{S}^1}(\varphi_a^*)$ are related by*

$$\Phi^*(\varphi_a^*) = \varphi_a/\Phi' \quad \textit{and} \quad \mathcal{F}^c_{\partial D}(\varphi_a) = M_{|\Phi'|^2}\big(\Phi^*(\mathcal{F}^c_{\mathbf{S}^1}(\varphi_a^*))\big).$$

Proof. If

$$\gamma : [0, L) \to \partial D$$

denotes a parametrization of ∂D by the arclength with positive orientation, then

$$\Phi \circ \gamma : [0, L) \to \mathbf{S}^1$$

is a regular $C^{1,\alpha}$-parametrization of $\mathbf{S}^1$ with positive orientation. Let $z \in \partial D$, $z = \gamma(t)$, and $w = \Phi(z) \in \mathbf{S}^1$. Then from $(\Phi \circ \gamma)^{\cdot}(t) = \Phi'(z) \cdot \dot{\gamma}(t)$ and $\dot{\gamma}(t) = \tau_D(z)$ we obtain that

$$\tau_{\mathcal{D}}(w) = \tfrac{\Phi'(z)}{|\Phi'(z)|}\tau_D(z) \quad \text{and} \quad \nu_{\mathcal{D}}(w) = \tfrac{\Phi'(z)}{|\Phi'(z)|}\nu_D(z). \tag{3.1}$$

The gradients of u and u^* are related by

$$\nabla u^*(w) = \nabla(u \circ \Phi^{-1})(w) = \nabla u(z) \circ J\Phi^{-1}(w).$$

By the above remarks the multiplication by $J\Phi^{-1}(w)$ corresponds to the complex multiplication by $\overline{(\Phi^{-1})'(w)} = 1/\overline{\Phi'(z)}$, hence

$$f^*(w) = f(z)/\Phi'(z). \tag{3.2}$$

In two dimensions the refraction operator can be written $R_a^{\partial D}(X) = \langle X, \tau\rangle\tau + a\langle X, \nu\rangle\nu$, $X \in \mathfrak{X}(\partial D)$. Rewriting the scalar products in complex notation, we obtain

$$\varphi_a^*(w) = \operatorname{Re}\big(f^*(w)\tau_{\mathcal{D}}(w)\big)\overline{\tau_{\mathcal{D}}(w)} + a\operatorname{Re}\big(f^*(w)\nu_{\mathcal{D}}\big)\overline{\nu_{\mathcal{D}}(w)}.$$

From this, inserting (3.1) and (3.2), we derive

$$\begin{aligned}\varphi_a^*(w) &= \operatorname{Re}\Big(\tfrac{f(z)\tau_D(z)}{|\Phi'(z)|}\Big)\overline{\Big(\tfrac{\Phi'(z)}{|\Phi'(z)|}\,\tau_D(z)\Big)} + a\operatorname{Re}\Big(\tfrac{f(z)\nu_D(z)}{|\Phi'(z)|}\Big)\overline{\Big(\tfrac{\Phi'(z)}{|\Phi'(z)|}\,\nu_D(z)\Big)}\\ &= \tfrac{\overline{\Phi'(z)}}{|\Phi'(z)|^2}\big\{\operatorname{Re}\big(f(z)\tau_D(z)\big)\cdot\overline{\tau_D(z)} + a\operatorname{Re}\big(f(z)\nu_D(z)\big)\cdot\overline{\nu_D(z)}\,\big\}\\ &= \varphi_a(z)/\Phi'(z).\end{aligned}$$

This, in view of $\Phi^*(\varphi_a^*)(z) = \varphi_a^*(w)$, proves the first statement of the lemma.

For the second statement let $h \in \mathcal{F}^c_{\mathbf{S}^1}(\varphi_a^*)$. By definition there exists $g \in \mathcal{H}^\alpha(\mathcal{D})$ such that

$$\operatorname{Re}(\overline{\varphi_a^*}g) = h \quad \text{on } \mathbf{S}^1.$$

Then the boundary values of $\hat{g} := \Phi' \cdot (g \circ \Phi) \in \mathcal{H}^\alpha(D)$ satisfy

$$\begin{aligned}\operatorname{Re}\big(\overline{\varphi_a(z)}\hat{g}(z)\big) &= \operatorname{Re}\big(\overline{\varphi_a^*(w)\Phi'(z)}\Phi'(z)g(w)\big)\\ &= |\Phi'(z)|^2 h(w) = |\Phi'(z)|^2\Phi^*(h)(z),\end{aligned}$$

whence $M_{|\Phi'|^2}\big(\Phi^*(\mathcal{F}^c_{\mathbf{S}^1}(\varphi_a^*))\big) \subset \mathcal{F}^c_{\partial D}(\varphi_a)$. The converse inclusion is derived in an analogous manner, by replacing Φ by its inverse Φ^{-1}. □

In particular, the mappings Φ^* and $M_{|\Phi|^2}$ are continuous and continuously invertible. Hence they map dense subsets of $L^1(\mathbf{S}^1)$ and $L^1(\partial D)$ onto dense subsets of $L^1(\partial D)$. Therefore the $L^1(\partial D)$-denseness of $\mathcal{F}^c_{\partial D}(\varphi_a)$ is equivalent to the $L^1(\mathbf{S}^1)$-denseness of $\mathcal{F}^c_{\mathbf{S}^1}(\varphi_a^*)$. Consequently, without loss of generality from now on we may assume that $D = \mathcal{D}$.

LEMMA 3.2. *The set of zeros* $\mathrm{N}(\varphi_1)$ *is closed and has measure zero in* $\mathbf{S}^1$. *The linear span of* $\mathcal{I}(\varphi_a)$ *is dense in* $L^p(\mathbf{S}^1)$ *for any* $1 \leqslant p \leqslant 2$.

Proof. The first statement is trivial. Proof that the set $\{z \in \mathbf{S}^1 : f(z) = 0\}$ has measure zero in $\mathbf{S}^1$ can be found in [12, Theorem 1.9, p. 203].

Since the $L^p(\mathbf{S}^1)$-norms, $p < 2$, are weaker than the $L^2(\mathbf{S}^1)$-norm, it suffices to show $L^2(\mathbf{S}^1)$-denseness of $\mathcal{I}(\varphi_a)$. Let $h' \in L^2(\mathbf{S}^1)$ be orthogonal to $\mathcal{I}(\varphi_a)$. Then for any $z_0 \in \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ we choose a closed and connected neighborhood $J \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ of z_0 and a function

$$\chi \in C^\alpha(\mathbf{S}^1, J) \quad \text{such that} \quad \chi \geqslant 0, \quad \chi(z_0) > 0.$$

We select a sequence $(h_k)_{k\in\mathbb{N}}$ in $C^\alpha(\mathbf{S}^1)$ converging to h' in $L^2(\mathbf{S}^1)$. Then, since $\chi h_k \in \mathcal{I}(\varphi_a)$, we have

$$\int_{\mathbf{S}^1} h'\chi h_k\, ds = 0 \quad \text{for all } k \in \mathbb{N}.$$

Passing to the limit $k \to \infty$ we find $\int_{\mathbf{S}^1}\chi(h')^2\, ds = 0$. Therefore h' vanishes almost everywhere in the neighborhood $\{z \in \mathbf{S}^1 : \chi(z) > 0\}$ of z_0. Since $\mathrm{N}(\varphi_a)$ has measure

zero, this implies $h' = 0$ almost everywhere. Thus $\mathcal{I}(\varphi_a)^\perp = \{0\}$, and the third assertion is established. □

LEMMA 3.3. *Let* arg *be the argument function on* $\mathbb{C} \setminus \{0\}$ *with values in* $(-\pi, \pi]$. *Then for all* $z \in \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$

$$\min(1, a) \leqslant |\varphi_a(z)/\varphi_1(z)| \leqslant \max(1, a), \tag{3.3}$$

$$|\arg(\varphi_a(z)/\varphi_1(z))| \leqslant \arccos\left(2\sqrt{a}/(1+a)\right) < \frac{\pi}{2}. \tag{3.4}$$

Proof. If $\nu(z)$, $\tau(z)$ is used as a coordinate system, then we have $\varphi_a(z) = (ax_1, x_2)$ provided that $\varphi_1(z) = (x_1, x_2)$. From this it follows that

$$\begin{aligned} |\varphi_1(z)| \leqslant |\varphi_a(z)| \leqslant a|\varphi_1(z)| \qquad & \text{if } a \geqslant 1, \\ a|\varphi_1(z)| \leqslant |\varphi_a(z)| \leqslant |\varphi_1(z)| \qquad & \text{if } a \leqslant 1, \end{aligned}$$

which establishes the first assertion. Interpreting $\nu(z)$ and $\tau(z)$ as complex numbers we define

$$\zeta_a(x_1, x_2) := \frac{ax_1\nu(z) + x_2\tau(z)}{x_1\nu(z) + x_2\tau(z)}, \quad (x_1, x_2) \in \mathbb{R}^2 \setminus \{0\},$$

and

$$\vartheta_a^* := \sup_{(x_1,x_2)\in\mathbb{R}^2\setminus\{0\}} |\arg(\zeta_a(x_1, x_2))|. \tag{3.5}$$

From these definitions it is obvious that $\left|\arg\left(\varphi_a(z)/\varphi_1(z)\right)\right| \leqslant \vartheta_a^*$. A simple computation shows that

$$\mathrm{Re}(\zeta_a(x_1, x_2)) = \frac{ax_1^2 + x_2^2}{x_1^2 + x_2^2} > 0, \quad (x_1, x_2) \in \mathbb{R}^2 \setminus \{0\}. \tag{3.6}$$

Since the function arg has discontinuities only on the half-line $\{z \in \mathbb{R} : z \leqslant 0\}$, the function $|\arg \circ \zeta_a|$ is continuous on $\mathbb{R}^2 \setminus \{0\}$. Since from the definitions it is obvious that ζ_a does depend only on the argument of (x_1, x_2), it suffices to take the supremum in (3.5) over the unit circle. Since continuous functions on compact sets assume their maximum, there exists (x_1^*, x_2^*) on $\mathbf{S}^1$ such that

$$\vartheta_a^* = |\arg(\vartheta_a(x_1^*, x_2^*))|.$$

From (3.6) we obtain $\vartheta_a^* < \frac{\pi}{2}$. Finally, the explicit expression

$$\vartheta_a^* = \arccos(2\sqrt{a}/(1+a))$$

follows from elementary calculations. □

LEMMA 3.4. *Suppose that* f *admits an extension as a holomorphic function in some neighborhood of* $\overline{\mathcal{D}}$. *Then the set* $\mathrm{N}(\varphi_1)$ *is finite and the function* $\varphi_a(z)/\varphi_1(z)$, $z \in \mathbf{S}^1 \setminus \mathrm{N}(\varphi_1)$ *can be extended as an analytic function* ψ *into* $\mathrm{N}(\varphi_1)$. *The extension* ψ *has no zeros and its index is zero.*

Proof. The smoothness of ψ need only be shown for points in the finite set $\mathrm{N}(\varphi_1)$. Let $z_0 \in \mathbf{S}^1$ with $\varphi_1(z) = 0$ and let u be the harmonic function with $\nabla u \simeq \overline{f}$. We claim that there exists a holomorphic function f_a defined in some ε-neighborhood $U_\varepsilon(z_0) := \{z \in \mathbb{C} : |z - z_0| < \varepsilon\}$ of z_0 such that

$$f_a|_{\mathbf{S}^1\cap U_\varepsilon} = \varphi_a. \tag{3.7}$$

Indeed, if ε is small enough, by the Cauchy–Kowalevski theorem there exists a solution v of

$$\triangle v = 0 \quad \text{in } U_\varepsilon(z_0) \qquad \text{and} \qquad v = u,\ \frac{\partial v}{\partial \nu} = a\frac{\partial u}{\partial \nu} \quad \text{on } \mathbf{S}^1 \cap U_\varepsilon(z_0).$$

Then obviously $\nabla v = R_a(\nabla u)$, and therefore the holomorphic function f_a defined by $\overline{f}_a \simeq \nabla v$ satisfies (3.7).

Thus in some neighborhood of z_0 on $\mathbf{S}^1$ the function φ_a/φ_1 is equal to the restriction of the meromorphic function f_a/f. If this function would have a pol in z_0, the values of $f_a(z)/f(z)$, $z \neq z_0$, would be unbounded in any neighborhood of z_0 in $\mathbf{S}^1$, and this contradicts (3.3) of Lemma 3.3. Consequently, f_a/f is holomorphic in some neighborhood of z_0, and this proves that φ_a/φ_1 admits an analytic extension onto all of $\mathbf{S}^1$.

Clearly, the estimates of Lemma 3.3 are valid on all of $\mathbf{S}^1$. From this, the last statement of Lemma 3.4 is obvious. □

For the solution of the Hilbert problem (2.10), we will have to define a logarithm of the function $-\psi/\overline{\psi}$ on $\mathbf{S}^1$. At this point, in the solution theory of a general Hilbert problem the index of ψ is important. In our special case we can use the estimate (3.4) to define a logarithm of $-\psi/\overline{\psi}$ somewhat more directly. Therefore the index of ψ will implicitly enter only in the proof of Theorem 2.1. Let $\vartheta_a^* := \arccos(2\sqrt{a}/(1+a))$ as in the proof of Lemma 3.3. With $\vartheta_a := \pi/2 - \vartheta_a^* > 0$ we can rewrite (3.4) in the form

$$\arg(\psi) \in [-\pi/2 + \vartheta_a, \pi/2 - \vartheta_a],$$

whence

$$\arg(\psi/\overline{\psi}) = \arg(\psi^2/|\psi|^2) = \arg(\psi^2) \in [-\pi + 2\vartheta_a, \pi - 2\vartheta_a]$$

follows. Defining a second argument function Arg with values in $[0, 2\pi)$, we get

$$\operatorname{Arg}(-\psi/\overline{\psi}) \in [2\vartheta_a, 2\pi - 2\vartheta_a].$$

If Log denotes the branch of the logarithm corresponding to Arg, that is, $\operatorname{Log}(z) = \log|z| + i\operatorname{Arg}(z)$ where log stands for the usual real logarithm, then the function

$$\operatorname{Log}(-\psi/\overline{\psi}) : \mathbf{S}^1 \to \mathbb{C}$$

is analytic on $\mathbf{S}^1$. Since the values of $-\psi/\overline{\psi}$ are on the unit circle, the values of $\operatorname{Log}(-\psi/\overline{\psi})$ are purely imaginary. For later use we note

$$\tfrac{1}{i}\operatorname{Log}\big(-\psi(z)/\overline{\psi}(z)\big) \in [2\vartheta_a, 2\pi - 2\vartheta_a], \qquad z \in \mathbf{S}^1. \tag{3.8}$$

There exist explicit formulas for the solution of (2.10) in terms of the Cauchy integral operator, which is defined by

$$A(\phi)(z) := \frac{1}{\pi i}\int_{\mathbf{S}^1} \frac{\phi(\zeta)}{\zeta - z}\, d\zeta, \qquad z \in \mathbb{C}, \quad \phi \in C^\alpha(\mathbf{S}^1)_{\mathbb{C}}.$$

We collect some properties of A from [5].

The function $A(\phi)$ is holomorphic in $\mathbb{C} \setminus \mathbf{S}^1$. The functions $A(\phi)_+ := A(\phi)|_{\mathbb{C}\setminus\overline{\mathcal{D}}}$ and $A(\phi)_- := A(\phi)|_{\mathcal{D}}$ can be extended onto $\mathbf{S}^1$ as functions in $C^\alpha(\mathbb{C}\setminus\mathcal{D})$ and $C^\alpha(\overline{\mathcal{D}})$, respectively. We have $\lim_{|z|\to\infty} A(\phi)(z) = 0$ uniformly in all directions. For $z \in \mathbf{S}^1$

the integral $A(\phi)(z)$ exists as a Cauchy principal value and $A(\phi)|_{\mathbf{S}^1} \in C^\alpha(\mathbf{S}^1)$. To avoid ambiguities, we define

$$\mathbf{A}(\phi) := A(\phi)|_{\mathbf{S}^1}, \qquad \phi \in C^\alpha(\mathbf{S}^1)_{\mathbb{C}}.$$

The mappings

$$C^\alpha(\mathbf{S}^1)_{\mathbb{C}} \to C^\alpha(\mathbf{S}^1)_{\mathbb{C}},\, C^\alpha(\mathbb{C}\setminus\mathcal{D})_{\mathbb{C}},\, C^\alpha(\bar{\mathcal{D}})_{\mathbb{C}}, \quad \phi \mapsto \mathbf{A}(\phi),\, A(\phi)_+,\, A(\phi)_-$$

are continuous. Their values on $\mathbf{S}^1$ are related by the Sokhotski–Plemelj formulas

$$A(\phi)_- - A(\phi)_+ = 2\phi \quad \text{and} \quad A(\phi)_- + A(\phi)_+ = 2\mathbf{A}(\phi) \quad \text{on } \mathbf{S}^1.$$

Later in the proof we will also need the facts that $\mathbf{A}^2 = \mathrm{Id}_{C^\alpha(\mathbf{S}^1)_{\mathbb{C}}}$ and that $\mathbf{A}$ can be extended as a continuous linear operator $L^2(\mathbf{S}^1)_{\mathbb{C}} \to L^2(\mathbf{S}^1)_{\mathbb{C}}$.

THEOREM 3.5. *If f admits an extension as a holomorphic function in some neighborhood of $\bar{\mathcal{D}}$, then the Hilbert problem* (2.5) *is solvable for any inhomogeneity $h \in \mathcal{I}(\varphi_a)$, that is, $\mathcal{I}(\varphi_a) \subset \mathcal{F}^c_{\mathbf{S}^1}(\varphi_a)$. If ψ is as in Lemma* 3.4*, then a solution g is given by*

$$H(z) := \exp\big(\tfrac{1}{2}A(\mathrm{Log}(-\psi/\overline{\psi}))\big)(z), \qquad z \in \mathbb{C}, \tag{3.9a}$$

$$G(z) := \tfrac{1}{2}H(z)\cdot A\big(-h/(|\varphi_1|^2\psi H_+)\big)(z), \qquad z \in \mathbb{C}, \tag{3.9b}$$

$$\tilde{g}(z) := G_-(z) + \overline{G_+(1/\overline{z})}, \qquad z \in \mathcal{D}, \tag{3.9c}$$

$$g(z) := f(z)\tilde{g}(z), \qquad z \in \mathcal{D}. \tag{3.9d}$$

Proof. As described in the preceding section, $\tilde{g}$ is the solution of

$$\tilde{g} \in \mathcal{H}^\alpha(\mathcal{D}) \text{ with } \mathrm{Re}(\overline{\psi}\tilde{g}) = h/|\varphi_1|^2. \tag{3.10}$$

The formulas (3.9a)–(3.9c) summarize the solution of (3.10) by transforming it into a Riemann problem (see [6]).

The Sokhotski–Plemelj formulas yield

$$\tfrac{1}{2}A(\mathrm{Log}(-\psi/\overline{\psi}))_- - \tfrac{1}{2}A(\mathrm{Log}(-\psi/\overline{\psi}))_+ = \mathrm{Log}(-\psi/\overline{\psi}) \quad \text{on } \mathbf{S}^1.$$

Taking the exponential, we deduce that H is a solution of the homogeneous Riemann problem

$$\overline{\psi}H_- + \psi H_+ = 0 \quad \text{on } \mathbf{S}^1.$$

Since $\exp z \neq 0$ for all $z \in \mathbb{C}$ the boundary values H_+ have no zeros. For brevity we set $\delta := -h/(|\varphi_1|^2\psi H_+)$. We recall the fact that we have set $\delta(z) = 0$ if $z \notin \operatorname{supp} h$, implying that $\delta \in C^\alpha(\mathbf{S}^1)_{\mathbb{C}}$. On $\mathbf{S}^1$ we have

$$\begin{aligned}\overline{\psi}G_- + \psi G_+ &= \tfrac{1}{2}(\overline{\psi}H_-A(\delta)_- + \psi H_+A(\delta)_+) = \tfrac{1}{2}\psi H_+(-A(\delta)_- + A(\delta)_+)\\ &= -\psi H_+\delta = h/|\varphi_1|^2.\end{aligned}$$

From $1/z = \overline{z}$, $z \in \mathbf{S}^1$, we find $\overline{G_+(1/\overline{z})} = \overline{G_+(z)}$, $z \in \mathbf{S}^1$, whence

$$\begin{aligned}\mathrm{Re}(\overline{\psi}\tilde{g}) &= \mathrm{Re}(\overline{\psi}G_- + \overline{\psi G_+}) = \mathrm{Re}(\overline{\psi}G_- + \psi G_+)\\ &= h/|\varphi_1|^2\end{aligned}$$

follows. Therefore the boundary values of g satisfy

$$\mathrm{Re}(\overline{\varphi}_a g) = |\varphi_1|^2 \,\mathrm{Re}(\overline{\psi}\tilde{g}) = h. \qquad \square$$

Now we consider the case of a general $f \in \mathcal{H}^\alpha(\mathcal{D})$ and for $0 < r < 1$ define

$$f_r(z) = f(rz), \quad \varphi_1^{(r)} := f_r|_{\mathbf{S}^1}, \quad \varphi_a^{(r)} := \overline{R_a^{\mathbf{S}^1}(\overline{f}_r|_{\mathbf{S}^1})}, \quad \psi_r := \varphi_a^{(r)}/\varphi_1^{(r)}.$$

The functions f_r are holomorphic in a neighborhood of $\bar{\mathcal{D}}$ and hence the functions ψ_r are analytic on $\mathbf{S}^1$. The function H in (3.9a) does not depend on the inhomogeneity h. Let

$$H(r;z) := \exp(A(i\phi_r)(z)), \quad z \in \mathbb{C}, \qquad \phi_r := \tfrac{1}{2i}\,\mathrm{Log}(-\psi_r/\overline{\psi}_r)$$

be formed correspondingly for f_r. From (3.8) we have

$$\phi_r(z) \in [\vartheta_a, \pi - \vartheta_a], \qquad z \in \mathbf{S}^1, \quad 0 < r < 1. \tag{3.11}$$

In the investigation of the convergence of $H(r;\cdot)$ for $r \to 1$, we will frequently use the fact that for a compact set $M \subset \mathbb{C}$ the multiplication mapping

$$C^\alpha(M)_{\mathbb{C}} \times C^\alpha(M)_{\mathbb{C}} \to C^\alpha(M)_{\mathbb{C}}, \quad (f,g) \mapsto fg,$$

and the inverse mapping

$$\{f \in C^\alpha(M)_{\mathbb{C}} : f(z) \neq 0,\, z \in M\} \to C^\alpha(M)_{\mathbb{C}}, \quad f \mapsto 1/f,$$

are continuous. Further, if $U \subset \mathbb{C}$ is open, then $C^\alpha(M,U)_{\mathbb{C}} := \{f \in C^\alpha(M)_{\mathbb{C}} : f(M) \subset U\}$ is open in $C^\alpha(M)_{\mathbb{C}}$ and, for $F \in C^\alpha(U)_{\mathbb{C}}$, the mapping

$$C^\alpha(M,U)_{\mathbb{C}} \to C^\alpha(M)_{\mathbb{C}}, \quad f \mapsto F \circ f,$$

is continuous. The fact that the images of these mappings are in $C^\alpha(M)_{\mathbb{C}}$ was already used above.

LEMMA 3.6. *If $J \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ is closed and connected, then the one-sided boundary values $H(r;\cdot)_\pm|_J$ are convergent in $C^\alpha(J)_{\mathbb{C}}$. The limit functions have no zeros in J.*

Proof. Let $\tilde{J} \subset \mathrm{N}(\varphi_a)$ be closed and connected. Since $\varphi_1^{(r)}$, $\varphi_a^{(r)}$ converge in $C^\alpha(\mathbf{S}^1)_{\mathbb{C}}$ to φ_1 and φ_a, both of which have no zeros on $\tilde{J}$, it follows that

$$(\psi_r/\overline{\psi}_r)|_{\tilde{J}} = (\varphi_a^{(r)}/\varphi_1^{(r)} \cdot \overline{\varphi_1^{(r)}}/\overline{\varphi_a^{(r)}})|_{\tilde{J}} \quad \text{converges for } r \to 1 \text{ in } C^\alpha(\tilde{J})_{\mathbb{C}}.$$

Hence

$$\phi_r|_{\tilde{J}} \text{ converges for } r \to 1 \text{ in } C^\alpha(\tilde{J})_{\mathbb{C}}. \tag{3.12}$$

In particular, ϕ_r is pointwise convergent outside the set $\mathrm{N}(\varphi_a)$ of measure zero. From (3.11) and Lebesgue's convergence theorem we obtain that

$$\phi_r \text{ converges for } r \to 1 \text{ in any } L^p(\mathbf{S}^1)_{\mathbb{C}},\ 1 \leqslant p < \infty. \tag{3.13}$$

Now, if $J \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ is closed and connected, the $C^\alpha(J)_{\mathbb{C}}$-convergence of $\mathbf{A}(i\phi_r)|_J$ follows by a suitable splitting of the integral. For this we choose a closed

and connected set $\tilde{J} \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ containing J in the interior and a real-valued $\chi \in C^\alpha(\mathbf{S}^1)$ such that

$$\operatorname{supp}\chi \subset \tilde{J}, \quad \chi \geqslant 0, \quad \chi = 1 \text{ on a neighborhood of } J.$$

Then $\chi\phi_r$, and in turn $\mathbf{A}(i\chi\phi_r)$, is $C^\alpha(\mathbf{S}^1)_{\mathbb{C}}$ convergent as $r \to 1$. The second part of the density $(1-\chi)\phi_r$ is $L^1(\mathbf{S}^1)_{\mathbb{C}}$-convergent and equal to zero in a neighborhood of J, which is independent of r. Hence $\mathbf{A}(i(1-\chi)\phi_r)|_J$ converges together with all its derivatives. Therefore

$$\mathbf{A}(i\phi_r)|_J = \mathbf{A}(i\chi\phi_r)|_J + \mathbf{A}(i(1-\chi)\phi_r)|_J \text{ converges for } r \to 1 \text{ in } C^\alpha(J)_{\mathbb{C}}.$$

The Sokhotski–Plemelj formulas imply $A(i\phi_r)_\pm = \mathbf{A}(i\phi_r) \mp i\phi_r$ on $\mathbf{S}^1$, whence, in view of (3.12), the $C^\alpha(J)_{\mathbb{C}}$ convergence of $A(i\phi_r)_\pm|_J$ follows. Since the application of the exponential preserves this convergence, the lemma is proved. □

The crucial point in part (iii) of the sketch of the proof is to show the uniform $L^1(\mathbf{S}^1)_{\mathbb{C}}$-boundedness of $H(r;\cdot)_\pm$. If it were possible to extend $\mathbf{A}$ to an endomorphism of $L^\infty(\mathbf{S}^1)_{\mathbb{C}}$, this would follow immediately from (3.11). But it is known that for a continuous, but not Hölder-continuous, density ϕ the function $\mathbf{A}(\phi)$ might have unbounded singularities. So $\mathbf{A}(L^\infty(\mathbf{S}^1)_{\mathbb{C}})$ is not even contained in $L^\infty(\mathbf{S}^1)_{\mathbb{C}}$. From (3.13) the density ϕ_r, and hence $\mathbf{A}(i\phi_r)$ and $A(i\phi_r)_\pm$, is $L^2(\mathbf{S}^1)_{\mathbb{C}}$-convergent as $r \to 1$. Thus we need a statement that the singularities of $\lim_{r\to 1}\mathbf{A}(i\phi_r)$ in $\mathrm{N}(\varphi_a)$ are weak enough such that $\exp(\mathbf{A}(i\phi_r))$ is still uniformly $L^1(\mathbf{S}^1)_{\mathbb{C}}$-bounded. This is achieved with the aid of the next theorem, which is due to Smirnov ([10, Hilfssatz 2]).

THEOREM 3.7. *Assume that $\phi \in C^\alpha(\mathbf{S}^1)$ satisfies $\phi(z) \in [\vartheta, \pi - \vartheta]$ for all $z \in \mathbf{S}^1$ and some $\vartheta > 0$. Then*

$$\|\exp(\mathbf{A}(i\phi))\|_{L^1(\mathbf{S}^1)_{\mathbb{C}}} \leqslant e^{\|\mathrm{Re}(\mathbf{A}(i\phi))\|_{L^1(\mathbf{S}^1)}} / \sin\vartheta.$$

Here we have set $\|F\|_{L^1(\mathbf{S}^1)_{\mathbb{C}}} := \frac{1}{2\pi}\int_0^{2\pi} \big|F(e^{it})\big|\,dt$, $F \in L^1(\mathbf{S}^1)_{\mathbb{C}}$.

Proof. Clearly, it suffices to prove the estimate for $\|\exp(\tilde{\phi})\|_{L^1(\mathbf{S}^1)}$ with $\tilde{\phi} := \mathrm{Re}(\mathbf{A}(i\phi))$. In the sequel we identify $C^\alpha(\mathbf{S}^1)_{\mathbb{C}}$ with $C^\alpha(\mathbb{R}/2\pi)_{\mathbb{C}}$ by the parametrization $t \mapsto e^{it}$.

On the unit circle the Cauchy integral operator can be written in the form

$$\begin{aligned}\mathbf{A}(\delta)(z) &= \frac{1}{\pi i}\int_{\mathbf{S}^1} \frac{\delta(\zeta)}{\zeta - z}\,d\zeta \\ &= \frac{1}{2\pi i}\int_0^{2\pi} \Big(\cot\frac{\tau - t}{2} + i\Big)\delta(\tau)\,d\tau, \qquad \delta \in C^\alpha(\mathbf{S}^1)_{\mathbb{C}}, \quad z = e^{it}.\end{aligned}$$

Denoting by $\mathbf{1}_{\mathbf{S}^1}$ the density equal to 1 everywhere on $\mathbf{S}^1$, it follows that $\mathbf{A}(i\phi) = \tilde{\phi} + ic_\phi\mathbf{1}_{\mathbf{S}^1}$, where $c_\phi = \frac{1}{2\pi}\int_0^{2\pi}\phi(\tau)\,d\tau$. Together with the formulas $\mathbf{A}^2 = \mathrm{Id}_{\mathbf{S}^1}$, $\mathbf{A}(\mathbf{1}_{\mathbf{S}^1}) = \mathbf{1}_{\mathbf{S}^1}$, and $\tilde{\phi} = \mathbf{A}(i\phi) - i\,\mathrm{Im}(\mathbf{A}(i\phi)) = \mathbf{A}(i\phi) - ic_\phi\mathbf{1}_{\mathbf{S}^1}$, this implies that

$$\begin{aligned}\mathbf{A}(\tilde{\phi} + i\phi) &= \mathbf{A}(\mathbf{A}(i\phi) - ic_\phi\mathbf{1}_{\mathbf{S}^1}) + \tilde{\phi} + ic_\phi\mathbf{1}_{\mathbf{S}^1} \\ &= i\phi + \tilde{\phi}.\end{aligned}$$

Now, from the Sokhotski–Plemelj formulas it can be seen that the holomorphic function

$$g(z) := \frac{1}{2\pi i}\int_{\mathbf{S}^1}\frac{\tilde{\phi}(\zeta) + i\phi(\zeta)}{\zeta - z}\,d\zeta = \tfrac{1}{2}\mathbf{A}(\tilde{\phi} + i\phi)(z), \quad z \in \mathcal{D},$$

has boundary values $g_- = \tilde{\phi} + i\phi$. Hence $e^g \in \mathcal{H}^\alpha(\mathcal{D})$ has boundary values $e^{\tilde{\phi}+i\phi}$. The Cauchy integral formula yields

$$e^{g(0)} = \frac{1}{2\pi i} \int_{\mathbf{S}^1} e^{(\tilde{\phi}+i\phi)(\zeta)} \, \frac{d\zeta}{\zeta} = \frac{1}{2\pi} \int_0^{2\pi} e^{\tilde{\phi}(\tau)+i\phi(\tau)} \, d\tau. \tag{3.14}$$

Taking the imaginary part of this equation we arrive at

$$\operatorname{Im}(e^{g(0)}) = \frac{1}{2\pi} \int_0^{2\pi} e^{\tilde{\phi}(\tau)} \sin(\phi(\tau)) \, d\tau.$$

Clearly $e^{\tilde{\phi}(\tau)} > 0$, and the assumption $\phi(\tau) \in [\vartheta, \pi - \vartheta]$ implies $\sin(\phi(\tau)) \geqslant \sin\vartheta > 0$. Thus

$$\operatorname{Im}(e^{g(0)}) / \sin\vartheta \geqslant \frac{1}{2\pi} \int_0^{2\pi} e^{\tilde{\phi}(\tau)} \, d\tau = \|\exp\tilde{\phi}\|_{L^1(\mathbf{S}^1)}.$$

If to the left-hand side we apply the estimate

$$\operatorname{Im}(e^{g(0)}) \leqslant \left| e^{g(0)} \right| = e^{\operatorname{Re} g(0)} = \exp\Big(\frac{1}{2\pi} \int_0^{2\pi} \tilde{\phi}(\tau) \, d\tau \Big) \leqslant e^{\|\tilde{\phi}\|_{L^1(\mathbf{S}^1)}},$$

the claim of the theorem follows. □

The asserted boundedness of the $H(r;\cdot)_\pm$ can now be proved.

COROLLARY 3.8. *The boundary values of the functions $H(r;\cdot)_\pm$ are uniformly $L^1(\mathbf{S}^1)_\mathbb{C}$-bounded as $r \to 1$.*

Proof. From Theorem 3.7 together with (3.11) we obtain that

$$\|H(r;\cdot)\|_{L^1(\mathbf{S}^1)_\mathbb{C}} \leqslant e^{\|\operatorname{Re}(\mathbf{A}(i\phi_r))\|_{L^1(\mathbf{S}^1)}} / \sin\vartheta_a. \tag{3.15}$$

In (3.12) we have already seen that ϕ_r and, consequently, $\mathbf{A}(i\phi_r)$ converge as $r \to 1$ in $L^2(\mathbf{S}^1)_\mathbb{C}$. Since on compact sets L^2-convergence is stronger than L^1-convergence, the right-hand side in (3.15) is convergent for $r \to 1$. Hence $H(r;\cdot)|_{\mathbf{S}^1}$ is uniformly bounded as $r \to 1$. Since the ϕ_r are real-valued, the assertion follows from

$$H(r;\cdot)_\pm = \exp(\mathbf{A}(i\phi_r) \mp i\phi_r) = H(r;\cdot)|_{\mathbf{S}^1} e^{\mp i\phi_r} \qquad \text{on } \mathbf{S}^1. \qquad \square$$

Now suppose we are given $h \in \mathcal{I}(\varphi_a)$ and $r_0 < 1$ is big enough to ensure that $h \in \mathcal{I}(\varphi_a^{(r)})$ for $r_0 < r < 1$. Then define functions $G(r;\cdot)$, $\tilde{g}_r$, and g_r by replacing φ_1, φ_a in the definitions (3.9b)–(3.9d) by $\varphi_1^{(r)}$ and $\varphi_a^{(r)}$, respectively, so that $g_r \in \mathcal{H}^\alpha(\mathcal{D})$ satisfy $\operatorname{Re}(\overline{\varphi_a^{(r)}} g_r) = h$ on $\mathbf{S}^1$. In view of the remarks at the end of (iii) in the preceding section, the proof of Theorem 2.1 is finished by the following lemma.

LEMMA 3.9. *The functions $g_r|_{\mathbf{S}^1}$ are uniformly $L^1(\mathbf{S}^1)_\mathbb{C}$-bounded for $r \to 1$.*

Proof. Let $J \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ be a closed and connected set such that $\operatorname{supp} h \subset J$. Consider the density

$$\delta_r := -h / \big(|\varphi_1^{(r)}|^2 \psi_r H(r;\cdot)_+ \big) = -h / \Big(\varphi_a^{(r)} \overline{\varphi_1^{(r)}} H(r:\cdot)_+ \Big),$$

of $\mathbf{A}$ in the definition (3.9b) of $G(r;\cdot)$. If $\tilde{J} \subset \mathbf{S}^1 \setminus \mathrm{N}(\varphi_a)$ is closed and connected and contains J in its interior, then by Lemma 3.6 the functions $\varphi_a^{(r)} \overline{\varphi_1^{(r)}} H(r;\cdot)_+|_{\tilde{J}}$ converge in $C^\alpha(\tilde{J})$ to some function without zeros in $\tilde{J}$. Since we have defined δ_r to

be zero outside the support of h, the density δ_r is $C^\alpha(\mathbf{S}^1)_{\mathbb{C}}$-convergent. Thus $\mathbf{A}(\delta_\mathbf{r})$ is convergent in $C^\alpha(\mathbf{S}^1)_{\mathbb{C}}$. The one-sided boundary values of $G(r;\cdot)$ are given by

$$G(r;z)_\pm = \tfrac{1}{2}H(r;z)_\pm \cdot (\mathbf{A}(\delta_r) \mp \delta_r(z)), \qquad z \in \mathbf{S}^1.$$

From

$$\|F_1F_2\|_{L^1(\mathbf{S}^1)_{\mathbb{C}}} \leqslant \|F_1\|_{L^1(\mathbf{S}^1)_{\mathbb{C}}}\|F_2\|_{L^\infty(\mathbf{S}^1)_{\mathbb{C}}}, \quad F_1 \in L^1(\mathbf{S}^1)_{\mathbb{C}},\ F_2 \in L^\infty(\mathbf{S}^1)_{\mathbb{C}},$$

the functions $G(r;\cdot)_\pm|_{\mathbf{S}^1}$ can be seen to be uniformly $L^1(\mathbf{S}^1)_{\mathbb{C}}$-bounded for $r \to 1$. Now it is easily checked from

$$\tilde{g}_r|_{\mathbf{S}^1} = G(r;\cdot)|_{\mathbf{S}^1} + \overline{G(r;\cdot)}|_{\mathbf{S}^1} \quad \text{and} \quad g_r = f_r\tilde{g}_r = \varphi_1^{(r)}\tilde{g}_r$$

that $\tilde{g}_r|_{\mathbf{S}^1}$ and $g_r|_{\mathbf{S}^1}$ are $L^1(\mathbf{S}^1)_{\mathbb{C}}$-bounded for $r \to 1$. □

Acknowledgments. I thank Professor Rainer Kress for his guidance during my research and his various suggestions for the improvement of this text.

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] H. Bellout and A. Friedman, *Identification problems in potential theory*, Arch. Rational Mech. Anal., 101 (1988), pp. 143–160.
[3] H. Bellout, A. Friedman, and V. Isakov, *Stability for an inverse problem in potential theory*, Trans. Amer. Math. Soc., 332 (1992), pp. 271–296.
[4] F. Hettlich and W. Rundell, *The determination of a discontinuity in a conductivity from a single boundary measurement*, Inverse Problems, 14 (1998), pp. 67–82.
[5] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.
[6] N. I. Muskhelishvili, *Singular Integral Equations*, P. Noordhoff, Groningen, the Netherlands, 1953.
[7] R. Potthast, *Fréchet differentiability of boundary integral operators and its application to an inverse accoustic scattering problem*, Inverse Problems, 10 (1994), pp. 431–447.
[8] R. Potthast, *Fréchet-Differenzierbarkeit von Randintegraloperatoren und Randwertproblemen zur Helmholtzgleichnung und den zeitharmonischen Maxwellgleichnungen*, Ph.D. thesis, Institut für Numerische und Angewandte Mathematik der Universität Göttingen, 1994.
[9] J. Powell, *On a small perturbation in the two-dimensional inverse conductivity problem*, J. Math. Anal. Appl., 175 (1993), pp. 292–304.
[10] V. Smirnov, *Über die Ränderzuordnung bei konformer Abbildung*, Math. Ann., 107 (1933), pp. 313–323.
[11] M. E. Taylor, *Partial Differential Equations* I*, Basic Theory*, Springer-Verlag, New York, 1996.
[12] A. Zygmund, *Trigonometric Series* II, 2nd ed., Cambridge University Press, Cambridge, UK, 1959.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 912–926

A STEFAN PROBLEM FOR A PROTOCELL MODEL*

AVNER FRIEDMAN† AND BEI HU‡

Abstract. The paper considers a simple model of a radially symmetric cell which undergoes growth due to a continuous supply of nutrient, and disintegration as a result of the various tasks the cell performs. The boundary of the cell is a "free boundary," unknown in advance, which evolves by responding to both the growth and disintegration processes. If the nutrient concentration (at infinity) exceeds a certain critical number, then two stationary solutions exist. It is established, by rigorous mathematical proofs, that the stationary solution with the smaller radius is unstable, whereas the stationary solution with the larger radius is stable.

Key words. protocell, parabolic system, free boundary problem

AMS subject classifications. 35K20, 35K55, 35K57, 35B40

PII. S0036141098337588

1. The model. In this paper we consider a physico-chemical model of a self-maintaining protocell which undergoes a process of growth and dissolution that mimics (but greatly simplifies) biological cells. The model is somewhat different from the one that was initiated and studied in [4], [5]. The protocell can be visualized as having a porous structure maintained by building materials with concentration C; the structure is sustained only as long as C exceeds a critical concentration C^*. Metabolism is maintained by nutrient material with concentration σ which is distributed in the entire space with $\sigma = \tau$ at ∞ ($\tau > 0$). C and σ satisfy a coupled system of reaction diffusion equations:

$$c\frac{\partial C}{\partial t} - \Delta C = \sigma, \quad \Delta\sigma - \sigma = 0 \quad \text{in the cell,}$$

$$\Delta\sigma = 0 \quad \text{outside the cell,}$$

where c is a positive constant. The constant c is the quotient of the time scale of diffusion to the time scale of cell doubling. In cases of interest, such as in tumor growth [1], [2], c is a very small constant.

On the boundary of the protocell $C = C^*$. The various tasks that the cell continuously performs take their toll on the cell: they cause it to shrink. This is modelled by disintegration at the boundary at a rate β, $\beta > 0$. On the other hand the flux of building material at the boundary causes the cell to grow. The total result of these two effects is

$$V_n = -\frac{\partial C}{\partial n} - \beta,$$

where n is the exterior normal, and V_n is the velocity of the boundary points in the direction n.

*Received by the editors April 17, 1998; accepted for publication October 14, 1998; published electronically June 4, 1999.
http://www.siam.org/journals/sima/30-4/33758.html
†IMA, University of Minnesota, Minneapolis, MN 55455 (friedman@ima.umn.edu). This author was partially supported by National Science Foundation grant DMS #9703842.
‡Department of Mathematics, University of Notre Dame, Notre Dame, IN 46556 (bei.hu.1@nd.edu). This author was partially supported by the Institute for Mathematics and Its Application during his visit there.

We shall consider here only the case of a spherical cell (in three dimensions). Setting

$$r = \sqrt{x_1^2 + x_2^2 + x_3^2}, \quad u = C - C^*$$

and denoting the boundary of the cell by $r = s(t)$, we then have the following system for $u = u(r,t), \sigma = \sigma(r,t)$ and $r = s(t)$:

$$\Big(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\Big)\sigma = \sigma\chi_{\{r<s(t)\}}(r,t) \quad \text{in } \mathbb{R}^3, \tag{1.1}$$

$$\sigma \to \tau \quad \text{as } |x| \to \infty, \tag{1.2}$$

and

$$cu_t - \Big(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\Big)u = \sigma \quad \text{if } r < s(t),\ t > 0, \tag{1.3}$$

$$u = 0 \quad \text{on } r = s(t),\ t > 0, \tag{1.4}$$

$$u = u_0(r) \quad \text{for } t = 0, \tag{1.5}$$

and finally, the free boundary condition

$$s'(t) = -u_r(s(t),t) - \beta. \tag{1.6}$$

As shown in [4], if τ/β is less than a critical number μ^*, then no steady state solutions exist, whereas if τ/β is larger than μ^*, then there exist two steady state solutions, with free boundary radii R_0^- and R_0^+, $R_0^- < R_0^+$. Numerical results and some heuristic arguments are given in [4] to show that the solution with R_0^- is unstable, whereas the solution with R_0^+ is stable. The purpose of this paper is to give rigorous mathematical proofs of these results.

In sections 2 and 3 we establish various estimates and prove the existence and uniqueness of the solution to (1.1)–(1.6). In section 4 we prove that the stationary solution corresponding to R_0^- is unstable. In section 5 we prove that the stationary solution corresponding to R_0^+ is stable if c is suitably small.

2. A priori bounds on the solution. The function

$$g(r) = \frac{r^2}{r - \tanh r} \quad (0 < r < \infty) \tag{2.1}$$

will play a fundamental role in what follows. One can easily verify that $r - \tanh r > 0$, so that $g(r) > 0$; furthermore, $g(r) \to \infty$ if $r \to 0$ or $r \to \infty$ and there is a unique $r = R^*$, where $g(r)$ attains its minimum μ^*, i.e.,

$$\min_{0<r<\infty} g(r) = g(R^*) = \mu^*. \tag{2.2}$$

One can compute that

$$R^* \approx 1.6061486 \quad \text{and} \quad \mu^* \approx 3.7739398.$$

The steady state solutions with free boundary $r = R_0$ are given by

$$u(r) = \frac{\tau}{\cosh R_0}\Big(\frac{\sinh R_0}{R_0} - \frac{\sinh r}{r}\Big) \quad \text{for } r < R_0, \tag{2.3}$$

where R_0 satisfies

$$\frac{\tau}{\beta} = \frac{R_0^2}{R_0 - \tanh R_0} = g(R_0). \tag{2.4}$$

If $\tau/\beta < \mu^*$, then there are no steady state solutions, whereas if $\tau/\beta > \mu^*$, then there are two solutions, $u_{R_0^+}$ and $u_{R_0^-}$, corresponding to $R_0 = R_0^+$ and $R_0 = R_0^-$, where $R_0^- < R^* < R_0^+$. As τ/β increases from μ^* to ∞, R_0^- decreases from R^* to 0 and R_0^+ increases from R^* to $+\infty$.

For a given $s(t)$, (1.1) with the boundary condition (1.2) can be solved explicitly (see [4, Eqs. (8) and (9)]):

$$\sigma(r,t) = \begin{cases} \tau\Big(1 - \dfrac{s(t) - \tanh s(t)}{r}\Big) & \text{for } r \geqslant s(t), \\ \tau\, \dfrac{1}{\cosh s(t)}\, \dfrac{\sinh r}{r} & \text{for } r < s(t). \end{cases} \tag{2.5}$$

Substituting (2.5) into (1.3), we get an equation for u involving the free boundary:

$$cu_t - \Big(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\Big)u = \tau\, \frac{1}{\cosh s(t)}\, \frac{\sinh r}{r} \quad \text{if } r < s(t),\ t > 0. \tag{2.6}$$

Further,

$$u = 0 \quad \text{on } r = s(t),\ t > 0, \tag{2.7}$$

$$s'(t) = -u_r(s(t),t) - \beta \quad \text{if } t > 0, \tag{2.8}$$

and

$$u = u_0(r) \quad \text{for } t = 0. \tag{2.9}$$

LEMMA 2.1. *If a solution of* (2.6)–(2.9) *exists for the time interval* $[0,T]$, *then*

$$\begin{aligned} \frac{4\pi}{3}s^3(t) = \frac{4\pi}{3}s^3(0) &+ c\int_{\{r<s(0)\}} u_0(r)dV - c\int_{\{r<s(t)\}} u(r,t)dV \\ &+ 4\pi\tau\int_0^t \Big[\frac{1}{g(s(\xi))} - \frac{\beta}{\tau}\Big]s^2(\xi)d\xi \end{aligned} \tag{2.10}$$

for $0 < t \leqslant T$.

Proof. By integration of (2.6) we get

$$\begin{aligned} &c\int_{\{r<s(t)\}} u(r,t)dV - c\int_{\{r<s(0)\}} u_0(r)dV \\ &= c\int_0^t \int_{\{r<s(\xi)\}} u_t(r,\xi)dV\,d\xi \\ &= \int_0^t \int_{\{r=s(\xi)\}} \frac{\partial u}{\partial r}dS + \tau\int_0^t \int_0^{s(\xi)} \frac{1}{\cosh s(t)}\frac{\sinh r}{r}4\pi r^2 dr\,d\xi \\ &= 4\pi\int_0^t \Big\{s^2(\xi)[-s'(\xi) - \beta] + \tau[s(\xi) - \tanh s(\xi)]\Big\}d\xi \\ &= \frac{4\pi}{3}[s^3(0) - s^3(t)] + 4\pi\int_0^t \{\tau[s(\xi) - \tanh s(\xi)] - \beta s^2(\xi)\}d\xi, \end{aligned}$$

from which the lemma follows. □

We shall henceforth assume that

$$0 \leqslant u_0(r) \leqslant C_0(s(0)-r) \quad \text{for } 0 \leqslant r < s(0), \tag{2.11}$$

where C_0 is a positive constant.

THEOREM 2.2. *If a solution* $(u(r,t), s(t))$ *of* (2.6)–(2.9) *exists for all* $0 < t < T$ $(T < \infty)$, *then*

$$s(t) < R \quad \textit{for all } 0 < t < T, \tag{2.12}$$

where R *is a constant independent of* T.

Proof. We distinguish the following two cases.

Case (i): $\tau/\beta \leqslant \mu^*$. Since $u_0(r) \geqslant 0$, we have, by the maximum principle, $u(r,t) > 0$ for $0 \leqslant r < s(t)$. From (2.10) and the definition of μ^* we obtain

$$\frac{4\pi}{3}s^3(t) \leqslant \frac{4\pi}{3}s^3(0) + c\int_{\{r<s(0)\}} u_0(r)dV - 4\pi\beta\int_0^t \Big(1 - \frac{1}{\mu^*}\frac{\tau}{\beta}\Big)s^2(\xi)d\xi. \tag{2.13}$$

It follows that

$$s(t) \leqslant \Big(s^3(0) + 3c\int_0^{s(0)} u_0(r)r^2dr\Big)^{1/3}. \tag{2.14}$$

Case (ii): $\tau/\beta > \mu^*$. In this case, the equation (2.3) has exactly two solutions R_0^- and R_0^+.

Setting

$$C_1 = \max(R_0^+, s(0)), \tag{2.15}$$

we can choose (using (2.11)) a constant R such that

$$C_1^3 + \frac{\tau}{2}cC_1^3R \leqslant R^3, \qquad \text{and } 0 \leqslant u_0(r) \leqslant \frac{\tau}{2}(R-r),\ 0 < r < s(0). \tag{2.16}$$

We shall prove that (2.12) holds for this choice of R. Indeed, if this is not true, then there is a first $t^* > 0$ such that $s(t^*) = R$. Since $s(t) < R$ for $0 < t < t^*$, we have, by comparison (using the maximum principle),

$$0 \leqslant u(r,t) \leqslant \frac{\tau}{2}(R-r), \quad 0 < r < s(t),\ 0 < t \leqslant t^*. \tag{2.17}$$

Take $t_1 \in [0, t^*)$ such that

$$s(t_1) = C_1, \quad C_1 < s(t) < R \quad \text{for } t_1 < t < t^*.$$

Then $\Big\{\frac{\tau}{\beta}[s(\xi) - \tanh s(\xi)] - s^2(\xi)\Big\} \leqslant 0$ for $t_1 < t < t^*$ and (2.10) yields

$$\begin{aligned} s^3(t) &< s^3(t_1) + 3c\int_0^{s(t_1)} u(r,t_1)r^2dr + 3\beta\int_{t_1}^t \Big\{\frac{\tau}{\beta}[s(\xi) - \tanh s(\xi)] - s^2(\xi)\Big\}d\xi \\ &\leqslant C_1^3 + \frac{\tau}{2}cC_1^3R \leqslant R^3 \quad \text{for } t_1 < t \leqslant t^*. \end{aligned}$$

Thus $s(t^*) < R$, which is a contradiction. □

From the uniform boundedness of $s(t)$ we can infer (by comparison) the uniform boundedness of $u(r,t)$:

$$0 \leqslant u(r,t) \leqslant C(R-r), \quad \text{where } R > \sup_{0\leqslant t\leqslant T} s(t). \tag{2.18}$$

The next lemma gives a sharper estimate on $u(r,t)$ for r near $s(t)$, as well as a useful estimate for $s'(t)$.

LEMMA 2.3. *Let R, C_1 be positive constant for which*

$$\begin{gathered} R > \sup_{0\leqslant t\leqslant T} s(t), \\ 0 \leqslant u_0(r) \leqslant C_1 \tanh(c\beta s(0)) \left\{1 - e^{-c\beta(s(0)-r)}\right\} \quad \textit{for } 0 \leqslant r < s(0). \end{gathered} \tag{2.19}$$

Then there is a constant K depending only on C_1, R, τ, c, and β such that

$$0 \leqslant u(r,t) \leqslant K \tanh(c\beta s(t)) \left\{1 - e^{-c\beta(s(t)-r)}\right\} \quad \textit{for } 0 < t \leqslant T, \tag{2.20}$$

and

$$-\beta < s'(t) < -\beta + cK \tanh(c\beta s(t)) \quad \textit{for } 0 < t \leqslant T. \tag{2.21}$$

Proof. By the maximum principle, $u(r,t) > 0$ if $r < s(t)$ and $u_r(s(t),t) < 0$; hence,

$$s'(t) > -\beta \quad \text{for } t > 0. \tag{2.22}$$

Let

$$w(r,t) = K \tanh(c\beta s(t)) \left\{1 - e^{-c\beta(s(t)-r)}\right\},$$

where we choose K large enough so that

$$w(r,0) \geqslant u_0(r). \tag{2.23}$$

Clearly,

$$\begin{aligned} & cw_t - \Big(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\Big) w \\ & = Kc\beta \tanh(c\beta s(t)) e^{-c\beta(s(t)-r)} \Big\{ cs'(t) + c\beta + \frac{2}{r} + \frac{c\left[e^{c\beta(s(t)-r)} - 1\right] s'(t)}{\sinh(c\beta s(t)) \cosh(c\beta s(t))} \Big\} \\ & \geqslant Kc\beta \tanh(c\beta s(t)) e^{-c\beta(s(t)-r)} \Big\{ \frac{2}{r} - \frac{2c\beta}{\cosh(c\beta s(t))} \Big\} \\ & \geqslant (2-\sqrt{2}) Kc\beta \tanh(c\beta s(t)) e^{-c\beta(s(t)-r)} \frac{1}{r}; \end{aligned}$$

in deriving the above inequalities, we made use of (2.22) and of the inequalities:

$$\begin{aligned} & e^{\xi} - 1 \leqslant 2\sinh\xi \quad \text{for } \xi \geqslant 0, \\ & \cosh(\xi) \geqslant 1 + \frac{\xi^2}{2} \geqslant \sqrt{2}\xi. \end{aligned}$$

We claim that for a suitable choice of the constant K,

$$(2-\sqrt{2})Kc\beta\tanh(c\beta s(t))e^{-c\beta(s(t)-r)}\frac{1}{r}\geqslant\tau\frac{1}{\cosh(s(t))}\frac{\sinh r}{r}\quad\text{for } 0<r<s(t); \tag{2.24}$$

the proof of this inequality will be given later. If (2.24) holds, then by the maximum principle,

$$u(r,t)\leqslant w(r,t). \tag{2.25}$$

(Note that $w_r|_{r=0}<0, u_r|_{r=0}=0$ so $w-u$ cannot take minimum at $r=0$.) It follows that

$$u_r(s(t),t)\geqslant w_r(s(t),t)=-Kc\beta\tanh(c\beta s(t)), \tag{2.26}$$

thus

$$s'(t)\leqslant-\beta+Kc\beta\tanh(c\beta s(t)).$$

To finish the proof of the lemma, it remains to verify (2.24). Observe that (2.24) is equivalent to

$$(2-\sqrt{2})K\tanh(c\beta s(t))\geqslant\frac{\tau}{c\beta}\,\frac{e^{c\beta s(t)}}{\cosh s(t)}e^{-c\beta r}\sinh r\quad\text{for } 0<r<s(t). \tag{2.27}$$

The function $e^{-c\beta r}\sinh r$ is monotonically increasing in the following cases:

Case (i). $c\beta\leqslant 1$;

Case (ii). $c\beta>1$, $r\leqslant s(t)\leqslant\frac{1}{2}\log\frac{c\beta+1}{c\beta-1}$ for all $0\leqslant t\leqslant T$.

In these two cases, it suffices to prove (2.27) for $r=s(t)$ so that (2.27) holds if K is such that

$$(2-\sqrt{2})K\geqslant\frac{\tau}{c\beta}\sup_{0<\xi<\infty}\frac{\tanh(\xi)}{\tanh(c\beta\xi)}\equiv\frac{\tau}{c\beta}\max\Big(1,\frac{1}{c\beta}\Big). \tag{2.28}$$

Finally, if $c\beta>1$ and $\max_{0\leqslant t\leqslant T}s(t)>\frac{1}{2}\log\frac{c\beta+1}{c\beta-1}$, we rewrite (2.24) in the form

$$(2-\sqrt{2})K\tanh(c\beta s(t))e^{-(c\beta-1)(s(t)-r)}\geqslant\frac{\tau}{c\beta}\,\frac{e^{s(t)}}{\cosh s(t)}e^{-r}\sinh r\quad\text{for } 0<r<s(t) \tag{2.29}$$

and observe that the function $e^{-r}\sinh r$ is monotonically increasing. It therefore suffices to prove (2.29) just for $r=s(t)$ on the right-hand side. But, since $s(t)<R$, this is a consequence of

$$(2-\sqrt{2})Ke^{-(c\beta-1)R}\geqslant\frac{\tau}{c\beta}\sup_{0<\xi<\infty}\frac{\tanh(\xi)}{\tanh(c\beta\xi)}\equiv\frac{\tau}{c\beta} \tag{2.30}$$

which holds if K is chosen sufficiently large. □

3. Existence and uniqueness. In this section we assume, in addition to (2.11), that

$$u_0'(r)\text{ is continuous if } 0\leqslant r\leqslant s(0). \tag{3.1}$$

THEOREM 3.1. *Let* (2.11) *and* (3.1) *hold. Then*

(i) *there exists a unique solution* $(u(r,t), s(t))$ *of* (2.6)–(2.9) *with* $s'(t)$ *continuous for a time interval* $0 \leqslant t < \delta_0$;

(ii) *if a solution* $(u(r,t), s(t))$ *exists for* $0 \leqslant t < T$ *and* $\liminf_{t \to T-0} s(t) > 0$, *then the solution can be continued to* $0 \leqslant t \leqslant T + \delta$ *for some* $\delta > 0$.

The proof, by a fixed point theorem for a contraction mapping, is similar to the corresponding proof for the Stefan problem [3, Chap. 8] and is therefore omitted.

THEOREM 3.2. *If*

$$\frac{\tau}{\beta} < \mu^*,$$

then there exists a finite t^* *such that the solution exists for* $0 \leqslant t < t^*$ *and* $s(t) \to 0$ *for* $t \to t^*$*; i.e., the cell shrinks to zero at time* t^*.

Proof. From (2.13) we get

$$s^3(t) + \delta \int_0^t s^2(\xi) d\xi \leqslant \widetilde{C}_1$$

for some positive constants $\widetilde{C}_1$ and δ.

By Theorem 3.1, the solution can be continued as long as $s(t)$ remains uniformly positive. If the assertion of the theorem is not true, then the solution exists (and $s(t)$ is positive) for all $0 < t < \infty$. Let K be as in Lemma 2.3 and take s_0 such that $K \tanh(\beta s_0) < \beta/2$. Since $\int_0^t s^2(\xi) d\xi$ is uniformly bounded, there exists a $\widetilde{t} > 0$ such that $s(\widetilde{t}) < s_0$. By Lemma 2.3 and a continuation argument, we then have $s'(t) \leqslant -\beta/2$ for all $t > \widetilde{t}$. This is a contradiction to the assumption that $s(t) > 0$ for all $t > 0$. □

In view of Theorems 2.2 and 3.1, a solution $(u(r,t), s(t))$ exists for all $0 \leqslant t < \infty$ if and only if, for any $T > 0$, there is $s > 0$ such that

$$\inf_{0 \leqslant t < T} s(t) \geqslant s > 0.$$

Furthermore, by Lemma 2.3, the constant s can be chosen to be independent of T if a global solution exists.

Next we establish a lower bound for $s(t)$ in the case $\tau/\beta > \mu^*$.

Let C_1, C_2 and D_1, D_2 be positive constants such that

$$C_1 \geqslant R_0^+, \quad C_1^3 + \frac{\tau}{2} c C_1^3 C_2 = C_2^3, \tag{3.2}$$

$$D_2 \geqslant R_0^-, \quad D_1 = D_2 \Big(\frac{\tau}{2} c C_2 + 1 \Big)^{1/3}. \tag{3.3}$$

THEOREM 3.3. *Assume that*

$$D_1 < R_0^+. \tag{3.4}$$

If $s(0), u_0(r)$ *satisfy*

$$D_1 \leqslant s(0) \leqslant C_1, \tag{3.5}$$

$$0 \leqslant u_0(r) \leqslant \frac{\tau}{2}(C_2 - r) \quad \text{for } 0 < r < s(0), \tag{3.6}$$

then there exists a global solution $(u(r,t), s(t))$ *of* (2.6)–(2.9) *with*

$$s(t) > D_2 \quad \text{for all } t > 0. \tag{3.7}$$

Proof. In view of (3.2), we can apply Case (ii) of Theorem 2.2 to conclude that $s(t) \leqslant C_2$ and then, by the maximum principle,

$$0 \leqslant u(r,t) \leqslant \frac{\tau}{2}(C_2 - r). \tag{3.8}$$

It follows that if

$$\begin{aligned} &R_0^- < s(t) < R_0^+ \quad \text{for } t_1 < t < t_2, \\ &s(t_1) = \min[s(0), R_0^+], \quad s(t_2) = D_2, \end{aligned}$$

then

$$3\int_0^{s(t_2)} u(r,t_2) r^2 dr - 3\int_0^{s(t_1)} u(r,t_1) r^2 dr \leqslant \frac{\tau}{2} C_2 D_2^3 = \frac{1}{c}(D_1^3 - D_2^3), \tag{3.9}$$

where the inequality follows from (3.8).

To prove the theorem it suffices to show that (3.7) holds as long as the solution exists. If this is not true then there is a smallest $t = t_2$ such that $s(t_2) = D_2$. By (3.4), (3.5) there exists a t_1 such that $t_1 < t_2$, $s(t_1) = \min[s(0), R_0^+]$ and $s(t_2) < s(t) < s(t_1)$ for $t_1 < t < t_2$. As in Case (ii) of Theorem 2.2,

$$s^3(t_1) - D_2^3 < 3c\int_0^{D_2} u(r,t_2) r^2 dr - 3c\int_0^{s(t_1)} u(r,t_1) r^2 dr,$$

and since $D_1^3 \leqslant \min[s^3(0), (R_0^+)^3]$, this is a contradiction to (3.9). □

Remark 3.1. If c is suitably small, then (3.4) is satisfied. In this case we have, by (3.7), (3.8),

$$D_2 < s(t) < C_2 \quad \text{for all } t > 0.$$

In particular, if

$$R_0^- + 2\delta < s(0) < C_1 \qquad (\delta > 0) \tag{3.10}$$

and c is sufficiently small, depending on δ and C_1, then

$$R_0^- + \delta < s(t) < C_1 + \delta \quad \text{for all } t > 0. \tag{3.11}$$

4. R_0^- is unstable. From Lemma 2.3 we infer that if $s(0)$ is very small, then $s'(t)$ remains uniformly negative as long as the solution exists. In this section we shall prove, under more general assumptions on $s(0)$, that $s(t)$ is monotone decreasing, and in particular, we shall deduce that the stationary solution corresponding to R_0^- is unstable.

THEOREM 4.1. *Suppose that*

$$u_r(r,0) \geqslant -\frac{\tau}{\cosh s(0)} \frac{r\cosh r - \sinh r}{r^2} - \frac{\tau}{s(0)}\Big[\frac{\beta}{\tau} - \frac{1}{g(s(0))}\Big] r, \tag{4.1}$$

$$u_r(s(0),0) > -\beta, \tag{4.2}$$

and

$$s(0) \leqslant R_0^-. \tag{4.3}$$

Then $s'(t) < 0$ as long as $s(t)$ remains positive, and $s(t)$ shrinks to zero in finite time.

Proof. From (4.2) we deduce that $s'(0) < 0$ and, by continuity, $s'(t) < 0$ for small $t > 0$. If $s'(t)$ does not remain negative while $s(t)$ is positive, then there exists a first t^* such that $s(t^*) > 0,\ s'(t^*) = 0$. Since $s'(t) < 0$ for $0 \leqslant t < t^*$, we have

$$\begin{aligned} &u_r(s(t), t) > -\beta \quad \text{for } 0 \leqslant t < t^*, \\ &0 < s(t) < s(0) \leqslant R_0^- \quad \text{for } 0 \leqslant t < t^*. \end{aligned}$$

We introduce the auxiliary function

$$w = u_r(r,t) + \frac{\tau}{\cosh s(t)} \frac{r\cosh r - \sinh r}{r^2} + \frac{\tau}{s(t)}\Big[\frac{\beta}{\tau} - \frac{1}{g(s(t))}\Big] r,$$

where $g(r)$ is defined in (2.1).

Since $g'(r) < 0$ for $0 < r \leqslant R_0^-$, $g(R_0^-) = \tau/\beta$, and $0 < s(t) \leqslant R_0^-$, we have

$$\Big[\frac{\beta}{\tau} - \frac{1}{g(s(t))}\Big] > 0 \quad \text{for } 0 < t < t^*. \tag{4.4}$$

Using the relation $(r\cosh r - \sinh r)/r^2 = (\partial/\partial r)(\sinh r/r)$, we easily deduce that

$$\begin{aligned} &-\Delta\Big(\frac{r\cosh r - \sinh r}{r^2}\Big) + \frac{2}{r^2}\Big(\frac{r\cosh r - \sinh r}{r^2}\Big) \\ &\quad = -\frac{\partial}{\partial r}\Delta\frac{\sinh r}{r} = -\frac{\partial}{\partial r}\frac{\sinh r}{r} = -\frac{r\cosh r - \sinh r}{r^2}. \end{aligned}$$

Combining this relation with (4.4) and the inequality $s'(t) < 0$, we find that

$$cw_t - w_{rr} - \frac{2}{r}w_r + \frac{2}{r^2}w > 0 \quad \text{for } 0 < t < t^*.$$

Clearly, $w(s(t), t) > 0$ for $0 < t < t^*$, and by (4.1), $w(r, 0) \geqslant 0$. It follows, by the maximum principle, that $w(r,t) > 0$ for $0 < t < t^*$. Since $s'(t^*) = 0$, we have also (from the definition of w and g) that $w(s(t^*), t^*) = 0$ and, therefore, by the maximum principle,

$$w_r(s(t^*), t^*) < 0.$$

That is,

$$u_{rr}(s(t^*), t^*) < -\frac{\tau}{s(t^*)}\Big[\frac{\beta}{\tau} - \frac{3}{s(t^*)} + \frac{3\tanh s(t^*)}{s^2(t^*)} + \tanh s(t^*)\Big].$$

On the other hand, from $s'(t^*) = 0$ we deduce that $u_t(s(t^*), t^*) = 0$ so that

$$\begin{aligned} 0 \;=\; cu_t(s(t^*), t^*) &= u_{rr}(s(t^*), t^*) + \frac{2}{s(t^*)}u_r(s(t^*), t^*) + \frac{\tau}{\cosh s(t^*)}\,\frac{\sinh s(t^*)}{s(t^*)} \\ &< -\frac{\tau}{s(t^*)}\Big[\frac{3\beta}{\tau} - \frac{3}{s(t^*)} + \frac{3\tanh s(t^*)}{s^2(t^*)}\Big] \\ &= -\frac{3\tau}{s(t^*)}\Big[\frac{\beta}{\tau} - \frac{1}{g(s(t^*))}\Big] \;\leqslant\; 0 \quad \text{by (4.4)}, \end{aligned}$$

which is a contradiction. Having now proved that $s'(t) < 0$ as long as $s(t)$ remains positive, we next show that $s(t)$ converges to zero in finite time. Indeed, otherwise the limit $s_0 = \lim_{t\to\infty} s(t)$ lies in the interval $(0, R_0^-)$. By a standard theorem on parabolic equations [3, Chap. 6], $(u(r,t), s(t))$ converges to a stationary solution, and consequently, s_0 must coincide with either R_0^- or R_0^+, which is a contradiction. □

If, in Theorem 4.1, we take $s(0) = R_0^-$, then we get the following.

COROLLARY 4.2. *Suppose*

$$u_r(r,0) \geqslant -\frac{\tau}{\cosh R_0^-}\frac{r\cosh r - \sinh r}{r^2} \quad \text{for } 0 < r < R_0^-, \tag{4.5}$$

$$s(0) = R_0^-, \qquad u_r(R_0^-, 0) > -\beta. \tag{4.6}$$

Then $s'(t) < 0$ as long as $s(t)$ remains positive, and $s(t)$ shrinks to zero in finite time.

Since there are arbitrarily small perturbations of the stationary solution

$$\frac{\tau}{\cosh R_0^-}\Big(\frac{\sinh R_0^-}{R_0^-} - \frac{\sinh r}{r}\Big)$$

which satisfies (4.5), (4.6), we conclude the following.

COROLLARY 4.3. *The stationary solution* (4.5) *is unstable.*

One can even choose small perturbations of the stationary solution with $s(0) > R_0^-$ (but $s(0) - R_0^-$ small) for which $s(t)$ shrinks to zero in finite time.

The method of proof of Theorem 4.1 can be extended to establish monotonic decrease of $s(t)$ under different assumptions on the data. We give one example.

THEOREM 4.4. *Suppose*

$$u_r(r,0) \geqslant -\beta\frac{r}{s(0)}, \qquad u_r(s(0),0) > -\beta, \tag{4.7}$$

$$\tau \leqslant \frac{3\beta}{\tanh s(0)}. \tag{4.8}$$

Then $s'(t) < 0$ as long as $s(t)$ remains positive.

The proof is essentially the same as the proof of Theorem 4.1, with w replaced by $u_r(r,t) + \beta r/s(t)$.

We note that Theorem 4.4 is not contained in Theorem 4.1 since, in general, $-\beta r/s(0)$ is not larger than the right-hand side of (4.1).

5. R_0^+ is stable. In this section we prove that the stationary solution corresponding to R_0^+ is stable provided the coefficient c is sufficiently small. As noted in the introduction, in actual biological cells of interest, c may indeed be very small. In the case $c = 0$, the solution $(\varphi(r,t), R(t))$ of (2.6)–(2.9) can be computed explicitly [4]:

$$\varphi(r,t) = \frac{\tau}{\cosh R(t)}\Big(\frac{\sinh R(t)}{R(t)} - \frac{\sinh r}{r}\Big) \quad \text{for } r < R(t),$$

where the free boundary $R(t)$ satisfies

$$\frac{dR}{dt} = \tau\frac{R - \tanh R}{R^2} - \beta = \tau\Big[\frac{1}{g(R)} - \frac{\beta}{\tau}\Big] \equiv \tau\Big(h(R) - \frac{\beta}{\tau}\Big). \tag{5.1}$$

Note that

$$h(R) > \frac{\beta}{\tau} \quad \text{and} \quad \dot{R} > 0 \quad \text{for } R_0^- < R < R_0^+,$$
$$h(R) < \frac{\beta}{\tau} \quad \text{and} \quad \dot{R} < 0 \quad \text{for } R > R_0^+,$$

and $h'(R_0^+) < 0$. Hence, by standard ODE analysis it follows that if $R(0) > R_0^-$, then

$$|R(t) - R_0^+| \leqslant Ce^{-\alpha t} \quad \text{for all } t > 0, \tag{5.2}$$

where C, α are positive constants.

In this section we want to extend this result to the case where c is positive and small.

THEOREM 5.1. *Let $s(0) > R_0^-$. If c is sufficiently small, then the solution $(u(r,t), s(t))$ of (2.6)–(2.9) exists for all $t > 0$,*

$$\lim_{t\to\infty} s(t) = R_0^+, \tag{5.3}$$

and the convergence is exponentially fast.

This establishes a global asymptotic stability of the stationary solution corresponding to R_0^+.

From Remark 3.1 we already know that the solution exists for all $t > 0$ and that (3.11) holds; i.e.,

$$R_0^- < \underline{S} \leqslant s(t) \leqslant \overline{S} < \infty \quad \text{for all } t > 0. \tag{5.4}$$

LEMMA 5.2. *Define*

$$\underline{R} = \liminf_{t\to\infty} s(t), \quad \overline{R} = \limsup_{t\to\infty} s(t).$$

If c is sufficiently small, then $\overline{R} = \underline{R} = R_0^+$.

Proof. By (5.4)

$$R_0^- < \underline{R} \leqslant \overline{R} \leqslant \overline{S}.$$

We claim that

$$R_0^- < \underline{R} \leqslant R_0^+ \leqslant \overline{R} \leqslant \overline{S}. \tag{5.5}$$

In fact, if $\underline{R} > R_0^+$, then $g(s(\xi)) > \tau/\beta + \varepsilon$ for some small $\varepsilon > 0$ and all sufficiently large ξ. Therefore,

$$\int_0^\infty \Big[\frac{1}{g(s(\xi))} - \frac{\beta}{\tau}\Big] s^2(\xi) d\xi = -\infty,$$

which is a contradiction to Lemma 2.1. This proves that $\underline{R} \leqslant R_0^+$. Similarly, $R_0^+ \leqslant \overline{R}$.

Next, we claim that

$$\liminf_{t\to\infty} u(r,t) \geqslant \frac{\tau}{\cosh \overline{R}} \Big(\frac{\sinh \underline{R}}{\underline{R}} - \frac{\sinh r}{r} \Big) \quad \text{uniformly for } r < s(t), \tag{5.6}$$

$$\limsup_{t\to\infty} u(r,t) \leqslant \frac{\tau}{\cosh \underline{R}} \Big(\frac{\sinh \overline{R}}{\overline{R}} - \frac{\sinh r}{r} \Big) \quad \text{uniformly for } r < s(t). \tag{5.7}$$

To prove (5.6), let $\varphi(r,t)$ be the solution of the following problem:

$$\begin{aligned} &c\varphi_t - \Delta\varphi = \frac{\tau}{\cosh(\overline{R}+\varepsilon)}\frac{\sinh r}{r} \quad \text{for } r < \underline{R}-\varepsilon,\ t > T,\\ &\varphi(\underline{R}-\varepsilon,t) = 0 \quad \text{for } t > T,\\ &\varphi(r,T) = 0 \quad \text{for } r < \underline{R}-\varepsilon, \end{aligned}$$

where $\varepsilon > 0$ is small. If we take $T = T(\varepsilon)$ to be large enough, then

$$\underline{R} - \varepsilon \leqslant s(t) \leqslant \overline{R} + \varepsilon \quad \text{for } t > T,$$

and, by maximum principle, $u(r,t) \geqslant \varphi(r,t)$ for $t > T$. Hence

$$\liminf_{t\to\infty} u(r,t) \geqslant \lim_{t\to\infty} \varphi(r,t) = \frac{\tau}{\cosh(\overline{R}+\varepsilon)}\Big(\frac{\sinh(\underline{R}-\varepsilon)}{\underline{R}-\varepsilon} - \frac{\sinh r}{r}\Big).$$

Letting $\varepsilon \to 0+$, we obtain the inequality (5.6). The inequality (5.7) can be established in a similar way.

Next, we estimate $\overline{R} - R_0^+$ in case $\overline{R} > R_0^+$. Choose a sequence $t_j \to \infty$ such that $s(t_j) \to \overline{R}$ and take $\widetilde{t}_j < t_j$ such that

$$R_0^+ + \frac{1}{j} = s(\widetilde{t}_j) < s(t) \quad \text{for } \widetilde{t}_j < t < t_j;$$

since $\liminf_{t\to\infty} s(t) \leqslant R_0^+ < \overline{R}$, such a choice of $\widetilde{t}_j$ is possible. Then $g(s(\xi)) > \tau/\beta$ for $\widetilde{t}_j < \xi < t_j$ and Lemma 2.1 implies that

$$s^3(t_j) - s^3(\widetilde{t}_j) \leqslant 3c\int_0^{s(\widetilde{t}_j)} u(r,\widetilde{t}_j)r^2dr - 3c\int_0^{s(t_j)} u(r,t_j)r^2dr.$$

Using (5.6), (5.7) in the last inequality, and letting $j \to \infty$, we obtain

$$\begin{aligned} \overline{R}^3 - (R_0^+)^3 &\leqslant 3c\int_0^{R_0^+} \frac{\tau}{\cosh\underline{R}}\Big(\frac{\sinh\overline{R}}{\overline{R}} - \frac{\sinh r}{r}\Big)r^2dr\\ &\quad -3c\int_0^{\underline{R}} \frac{\tau}{\cosh\overline{R}}\Big(\frac{\sinh\underline{R}}{\underline{R}} - \frac{\sinh r}{r}\Big)r^2dr\\ &\leqslant 3c\tau\int_0^{R_0^+} \Big(\frac{1}{\cosh\underline{R}}\frac{\sinh\overline{R}}{\overline{R}} - \frac{1}{\cosh\overline{R}}\frac{\sinh\underline{R}}{\underline{R}}\Big)r^2dr\\ &\quad +3c\tau\int_0^{R_0^+} \Big(\frac{1}{\cosh\overline{R}} - \frac{1}{\cosh\underline{R}}\Big)r\sinh r dr. \end{aligned}$$

Since the last integrand is $\leqslant 0$, we obtain

$$\begin{aligned} \overline{R}^3 - (R_0^+)^3 &\leqslant c\tau\Big(\frac{1}{\cosh\underline{R}}\frac{\sinh\overline{R}}{\overline{R}} - \frac{1}{\cosh\overline{R}}\frac{\sinh\underline{R}}{\underline{R}}\Big)(R_0^+)^3\\ &= \frac{c\tau}{\cosh\underline{R}\cosh\overline{R}}[k(\overline{R}) - k(\underline{R})](R_0^+)^3, \end{aligned} \tag{5.8}$$

where $k(r) = \sinh r\cosh r/r$; note that $k'(r) > 0$. Similarly, we have (with $s(\tau_j) \to \underline{R}, s(\widetilde{\tau}_j) = R_0^+ - 1/j, s(t) < R_0^+ - 1/j$ if $\widetilde{\tau}_j < t < \tau_j$)

$$\underline{R}^3 - (R_0^+)^3 \geqslant -\frac{c\tau}{\cosh\underline{R}\cosh\overline{R}}[k(\overline{R}) - k(\underline{R})](R_0^+)^3, \tag{5.9}$$

provided $\underline{R} < R_0^+$. Notice that if $\overline{R} = R_0^+$ (or $\underline{R} = R_0^+$) then (5.8) (or (5.9)) is trivially satisfied. Combining (5.8) with (5.9) we get

$$\overline{R} - \underline{R} \leqslant \frac{c\tau}{A}(\overline{R} - \underline{R}), \tag{5.10}$$

where

$$A = \frac{(\overline{R}^2 + \overline{R}\underline{R} + \underline{R}^2)\cosh \underline{R} \cosh \overline{R}}{2(R_0^+)^3 k'(\overline{R})} \geqslant A_0,$$

and A_0 is a positive constant depending only on $\underline{S}$, $\overline{S}$, and R_0^+. If $c\tau/A_0 < 1$, then $\overline{R} - \underline{R} = 0$ and the lemma is proved. □

Having proved (5.2), we shall next prove local stability of R_0^+ with exponential convergence of $s(t)$. Introduce the functions

$$\varphi(r,t) = \frac{\tau}{\cosh s(t)}\left(\frac{\sinh s(t)}{s(t)} - \frac{\sinh r}{r}\right), \tag{5.11}$$

and $v = u - \varphi$. A direct computation shows that

$$cv_t - \Big(\frac{\partial^2}{\partial r^2} + \frac{2}{r}\frac{\partial}{\partial r}\Big)v = -c\varphi_t \quad \text{for } 0 \leqslant r < s(t), \tag{5.12}$$

$$v(s(t),t) = 0, \tag{5.13}$$

$$s'(t) = -v_r(s(t),t) - \beta + \tau\frac{s(t) - \tanh s(t)}{s^2(t)}. \tag{5.14}$$

LEMMA 5.3. *Suppose that $c\tau < 3/2$. Then there exists an $\varepsilon > 0$ such that if*

$$|s(0) - R_0^+| < \varepsilon^2, \quad |s'(0)| < \varepsilon,$$

and

$$\left|u_0(r) - \frac{\tau}{\cosh s(0)}\Big(\frac{\sinh s(0)}{s(0)} - \frac{\sinh r}{r}\Big)\right| < \frac{\tau}{7}\varepsilon\Big(s(0) - \frac{r^2}{s(0)}\Big),$$

then a global solution $(u(r,t), s(t))$ exists and $s(t)$ converges to the steady state (corresponding to R_0^+) exponentially fast.

The proof requires the fact that $g'(R_0^+) > 0$ (g is defined in (2.1)) and, thus, does not apply to R_0^- (since $g'(R_0^-) < 0$).

Proof. Clearly,

$$\begin{aligned}
|\varphi_t| &= \tau|s'(t)|\left|\frac{s(t) - \tanh s(t)}{s^2(t)} - \frac{\sinh s(t)}{\cosh^2 s(t)}\Big(\frac{\sinh s(t)}{s(t)} - \frac{\sinh r}{r}\Big)\right| \\
&\leqslant \tau|s'(t)| \max\Big[\frac{s(t) - \tanh s(t)}{s^2(t)}, \frac{\sinh s(t)}{\cosh^2 s(t)}\Big(\frac{\sinh s(t)}{s(t)} - \frac{\sinh r}{r}\Big)\Big] \\
&\leqslant \tau|s'(t)| \max\Big[\frac{s(t) - \tanh s(t)}{s^2(t)}, \frac{\sinh^2 s(t)}{s(t)\cosh^2 s(t)}\Big] \\
&\leqslant \tau|s'(t)|\frac{1}{s(t)}.
\end{aligned} \tag{5.15}$$

We claim that

$$|s'(t)| < \varepsilon e^{-\varepsilon t} \quad \text{for } 0 < t < \infty. \tag{5.16}$$

By assumption, (5.16) is satisfied for $t = 0$. If the assertion (5.16) is not true, then there is a first t^* such that $|s'(t^*)| = \varepsilon e^{-\varepsilon t^*}$. Introduce the function

$$w(r,t) = \frac{c\tau\varepsilon}{6 - C^*\varepsilon} e^{-\varepsilon t}\Big(s(t) - \frac{r^2}{s(t)}\Big) \quad \text{for } 0 \leqslant t \leqslant t^*,$$

where $C^* > s^2(t) + 2c$ for all $0 < t < \infty$. Then

$$\begin{aligned} cw_t - \Delta w &= \frac{c\tau\varepsilon}{6 - C^*\varepsilon}\ \frac{1}{s(t)}\ \Big[6 + c\Big(1 + \frac{r^2}{s^2(t)}\Big)s'(t) - \varepsilon(s^2(t) - r^2)\Big]e^{-\varepsilon t} \\ &\geqslant \frac{c\tau\varepsilon}{6 - C^*\varepsilon}\ \frac{1}{s(t)}(6 - C^*\varepsilon)e^{-\varepsilon t} \\ &\geqslant \frac{c\tau\varepsilon}{s(t)} e^{-\varepsilon t}, \end{aligned}$$

and, by assumption, $w(r,0) > |v(r,0)|$. Recalling (5.15) and (5.12), (5.13), we can use the maximum principle to compare $\pm v$ with w and conclude that

$$|v(r,t)| \leqslant w(r,t).$$

Since $v(s(t),t) = w(s(t),t)$, we have also

$$|v_r(s(t),t)| \leqslant \frac{2c\tau\varepsilon}{6 - C^*\varepsilon} e^{-\varepsilon t}.$$

Thus

$$\Big|s'(t) + \beta - \tau\frac{s(t) - \tanh s(t)}{s^2(t)}\Big| \leqslant \frac{2c\tau\varepsilon}{6 - C^*\varepsilon} e^{-\varepsilon t}. \tag{5.17}$$

Let

$$h(r) = \frac{1}{g(r)}, \qquad f(t) = \tau\frac{h(R_0^+) - h(s(t))}{s(t) - R_0^+},$$

where $g(r)$ is defined by (2.1). Then we can rewrite (5.17) in the form

$$\Big|(s(t) - R_0^+)' + f(t)(s(t) - R_0^+)\Big| \leqslant \frac{2c\tau\varepsilon}{6 - C^*\varepsilon} e^{-\varepsilon t}. \tag{5.18}$$

It follows that

$$|s(t) - R_0^+| \leqslant |s(0) - R_0^+| \exp\Big(-\int_0^t f(\xi)d\xi\Big) + \frac{2c\tau\varepsilon}{6 - C^*\varepsilon}\int_0^t \exp\Big(-\varepsilon\eta - \int_\eta^t f(\xi)d\xi\Big)d\eta.$$

Notice that for $s(t)$ near R_0^+, $(h(R_0^+) - h(s(t)))/(s(t) - R_0^+) \sim -h'(R_0^+) \equiv \delta > 0$. Thus, for $s(t)$ near R_0^+, $\tau(\delta + C^*\varepsilon) \geqslant f(t) \geqslant \tau(\delta - C^*\varepsilon)$. Consequently,

$$|s(t) - R_0^+| \leqslant \frac{\varepsilon^2}{\tau(\delta - C^*\varepsilon)} e^{-\tau(\delta - C^*\varepsilon)t} + \frac{2c\tau\varepsilon}{6 - C^*\varepsilon}\ \frac{1}{\tau(\delta - C^*\varepsilon) - \varepsilon} e^{-\varepsilon t}. \tag{5.19}$$

Using this in (5.18) we obtain

$$\begin{aligned} |s'(t)| &\leqslant \frac{2c\tau\varepsilon}{6 - C^*\varepsilon} e^{-\varepsilon t} + |f(t)||s(t) - R_0^+| \\ &\leqslant \frac{2c\tau\varepsilon}{6 - C^*\varepsilon} e^{-\varepsilon t} \\ &\qquad + \tau(\delta + C^*\varepsilon)\left(\frac{\varepsilon^2}{\tau(\delta - C^*\varepsilon)} e^{-\tau(\delta - C^*\varepsilon)t} + \frac{2c\tau\varepsilon}{6 - C^*\varepsilon}\ \frac{1}{\tau(\delta - C^*\varepsilon) - \varepsilon} e^{-\varepsilon t}\right) \\ &< \varepsilon e^{-\varepsilon t} \quad \text{for } 0 < t \leqslant t^*, \end{aligned}$$

provided

$$\frac{2c\tau}{6-C^*\varepsilon}+\Big[\frac{\tau(\delta+C^*\varepsilon)\varepsilon}{\tau(\delta-C^*\varepsilon)}+\frac{2c\tau}{6-C^*\varepsilon}\,\frac{\tau(\delta+C^*\varepsilon)}{\tau(\delta-C^*\varepsilon)-\varepsilon}\Big]<1,$$

which is satisfied if $c\tau<3/2$ and ε is sufficiently small, and this is a contradiction to the assumption that $|s'(t^*)|$ is equal to $e^{-\varepsilon t^*}$. Finally, from (5.19), we deduce that $s(t)$ converges to R_0^+ exponentially fast. □

Proof of Theorem 5.1. From Lemma 5.2 and (5.6), (5.7), we see that the assumptions of Lemma 5.3 are satisfied at some sufficiently large time $t=T$. It follows that

$$|s(t)-R_0^+|\leqslant Ce^{-\alpha t},$$
$$|s'(t)|\leqslant Ce^{-\alpha t}$$

for all large enough t, where C,α are positive constants, and this completes the proof of Theorem 5.1. □

Remark 5.1. The above proof shows that if (3.6) and (3.10) hold, then the assertion of Theorem 5.1 holds for any $0<c<c^*$, where c^* depends only on δ, C_1, and C_2 in addition to τ and β.

REFERENCES

[1] H.M. Byrne and M.A.J. Chaplain, *Growth of nonnecrotic tumors in the presence of inhibitors*, Math. Biosci., 130 (1995), pp. 151–181.

[2] H.M. Byrne and M.A.J. Chaplain, *Growth of nonnecrotic tumors in the presence of inhibitors*, Math. Biosci., 135 (1996), pp. 187–216.

[3] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[4] H. Schwegler, K. Tarumi, and B. Gerstmann, *Physico-chemical model of protocell*, J. Math. Biol., 22 (1985), pp. 335–348.

[5] K. Tarumi and H. Schwegler, *A nonlinear treatment of the protocell by boundary layer approximation*, Bull. Math. Biol., 49 (1987), pp. 307–320.

SIAM J. MATH. ANAL.
Vol. 30, No. 4, pp. 927–936

AN ANALYTICAL PROOF OF THE LINEAR STABILITY OF THE VISCOUS SHOCK PROFILE OF THE BURGERS EQUATION WITH FOURTH-ORDER VISCOSITY*

SHLOMO ENGELBERG†

Abstract. In this paper we establish the exponential decay of solutions of the equation

$$u_t + \varphi(x)u_x = -\partial_x^4 u$$

in an exponentially weighted norm. Here $\varphi(x)$ is the viscous shock profile corresponding to the Burgers equation with fourth-order viscosity:

$$u_t + uu_x = -\partial_x^4 u.$$

Because of the fact that the profile is not monotone, showing the stability is nontrivial. We extend the techniques of Koppel and Howard (*Adv. Math.* 18 (1975), pp. 306-358), techniques that they employ to prove the existence of the viscous shock profile, and we use the techniques to prove the stability of the viscous shock profile. We have previously shown that the viscous shock profile is a stable solution in an exponentially weighted norm by making use of numerical results. The main advantage of our current method is that it is analytical. One sees more clearly what properties of the viscous shock profile cause it to be a stable solution of the PDE.

Key words. viscous shock profiles, stability

AMS subject classifications. 35B35, 35K55, 34E05

PII. S003614109833639X

1. Introduction.

In this paper we establish the exponential decay of solutions of the equation

$$u_t + \varphi(x)u_x = -\partial_x^4 u \tag{1.1}$$

in an exponentially weighted norm. Here $\varphi(x)$ is the stationary (i.e., time-independent) viscous shock profile corresponding to the Burgers equation with fourth-order viscosity,

$$u_t + (f(u))_x = -\partial_x^4 u, \quad f(u) = u^2/2, \tag{1.2}$$

that satisfies the condition $\varphi(x) \to \mp 1$ as $x \to \pm\infty$.

Equation (1.2) occurs in many contexts. Sivashinsky [10] has shown that (1.2) can be used to model burning on a Bunsen burner. Additionally, the modified equations that correspond to stable third-order methods for approximating the solutions of

$$u_t + uu_x = 0 \tag{1.3}$$

have the general form

$$u_t + uu_x = -k(u)u_{xxxx}. \tag{1.4}$$

*Received by the editors March 30, 1998; accepted for publication September 24, 1998; published electronically June 4, 1999. This work was performed while the author was with Tel Aviv University's School of Mathematical Sciences.

http://www.siam.org/journals/sima/30-4/33639.html

†Electronics Department, Jerusalem College of Technology—Machon Lev, P.O.B. 16031, Jerusalem 91160, Israel (shlomoe@optics.jct.ac.il).

We show that the stationary viscous profile of (1.4) is linearly stable when $k(u) \equiv 1$.

Furthermore, the problems that arise in the study of the fourth-order version of Burgers's equation are similar to the problems that come up in the study of the stability of viscous shock profiles of systems of equations. Work on the stability of viscous shock profiles of systems of equations has been done by Goodman [3], Liu [5], Szepessy and Xin [11], and Matsumura and Nishihara [6].

We have shown previously [2] that showing the exponential decay of solutions of (1.1) in a slightly different exponentially weighted norm is sufficient to show that the viscous shock profile is a nonlinearly stable solution of (1.2) in that exponentially weighted norm. It is simple to extend that proof to show that our results here are sufficient to guarantee that the solution is nonlinearly stable in our other, slightly different norm.

In [2], we made use of rigorous numerical results of Michelson [8] and the theory of sectorial operators to show that the solutions of (1.1) are exponentially stable in an exponentially weighted norm. Our current technique is completely analytical and does not use the theory of sectorial operators.

The function $\varphi(x)$ is the odd solution of the ODE $\varphi''' = \frac{1}{2}(1-\varphi^2)$ subject to the condition $\varphi(\infty) = -1$. Koppel and Howard [4] established the existence of the oddly symmetric viscous shock profile by showing that there exists a ν such that the solution of the ODE $\varphi''' = \frac{1}{2}(1-\varphi^2)$ with initial values $\varphi = 0, \varphi' = -\nu, \varphi'' = 0$ must intersect the stable manifold of the solution -1. Later, McKord [7] showed that this is the *unique* solution of the ODE subject to the conditions that the solution be odd and tend to -1 as $x \to \infty$. We show that zero is a stable solution of the linearization of the integrated Burgers equation about the viscous shock profile, i.e., of (1.1), by making use of some estimates contained in the existence proof given by Koppel and Howard [4].

As has been pointed out by Alexander, Gardner, and Jones [1], some existence proofs contain within them the seeds of stability proofs. In [1], there are results of this nature for systems of second-order PDEs. In this paper, we deal with solutions of a fourth-order equation in an exponentially weighted space.

Existence proofs for $\varphi(x)$ abound. The earliest proof is by Koppel and Howard [4]. It is a *tour de force*. Mock [9] presents a proof of the existence of viscous shock profiles for equations of the type we are studying. His proof is somewhat more general than that of Koppel and Howard. Finally, McKord [7] presents an index theoretic proof of the existence of the viscous shock profile. We show that Koppel and Howard's methods can be extended to a proof of stability in an exponentially weighted space.

As it seems that Koppel and Howard's general method can also be used to prove the existence of viscous shock profiles for Burgers's equation with more general fluxes than $f(u) = u^2/2$, it seems likely that our method can be used to prove the stability of such profiles in an exponentially weighted space. Then the methods of our previous article should suffice to show that the profiles are nonlinearly stable as well.

2. Conditions sufficient for decay to prevail. We look at the behavior of solutions of (1.1) in the w_ϵ norm which is defined as

$$\|g\|^2_{L^2_{w,\epsilon}} = \int_{-\infty}^{\infty} w_\epsilon(x) g^2(x)\, dx, \qquad w_\epsilon(x) = \cosh(\epsilon x).$$

We prove that the derivative of the w_ϵ norm of the solutions of (1.1) is negative.

Calculating the derivative and integrating by parts, we find that

$$\frac{d}{dt}\|u(t)\|^2_{L^2_{w,\epsilon}} = \epsilon \int \sinh(\epsilon x)\varphi(x)u^2(x)\,dx + \int w_\epsilon(x)\varphi'(x)u^2\,dx - 2\int (w_\epsilon u)_{xx}u_{xx}\,dx.$$

We note that if $\varphi'(x) \le 0$, then $\epsilon = 0$ is sufficient to force the norm of the solution not to increase. It is because our profile is not monotone that we must make use of a weighted norm in order to get any sort of linear stability result.

A simple calculation shows that

$$\int (w_\epsilon u)_{xx}u_{xx}\,dx = \int (\epsilon^2 w_\epsilon u + 2\epsilon \sinh(\epsilon x)u_x + w_\epsilon u_{xx})u_{xx}\,dx.$$

Integration by parts leads to

$$\int \epsilon^2 w_\epsilon u u_{xx}\,dx = \int \frac{\epsilon^4}{2} w_\epsilon u^2\,dx - \epsilon^2 \int w_\epsilon u_x^2\,dx,$$

and

$$\int 2\epsilon \sinh(\epsilon x)u_x u_{xx} = -\epsilon^2 \int w_\epsilon(x)u_x^2\,dx.$$

Finally, we find that

$$\begin{aligned}\frac{d}{dt}\|u(t)\|^2_{L^2_{w,\epsilon}} &= \epsilon \int \sinh(\epsilon x)\varphi(x)u^2(x)\,dx + \int w_\epsilon(x)\varphi'(x)u^2\,dx \\ &\quad -2\int w_\epsilon(x)(u_{xx})^2\,dx + 4\epsilon^2 \int w_\epsilon(x)u_x^2\,dx - \epsilon^4 \int w_\epsilon u^2\,dx.\end{aligned}$$

To finish off our calculations, we estimate these terms in such a way that they can be written as the integral of the product the weight, $w_\epsilon(x)$, a function to be determined, $\Psi(x)$, and $u^2(x)$. To do this we bound $\int w_\epsilon u_x^2\,dx$ as follows:

$$\begin{aligned}\int w_\epsilon u_x^2\,dx &= \frac{1}{2}\epsilon^2 \int w_\epsilon u^2\,dx - \int w_\epsilon u u_{xx}\,dx \\ &\le \frac{1}{2}\epsilon^2 \int w_\epsilon u^2\,dx + \sqrt{\int w_\epsilon u^2\,dx \int w_\epsilon u_{xx}^2\,dx} \\ &\le \frac{1}{2}\epsilon^2 \int w_\epsilon u^2\,dx + \frac{c}{2}\int w_\epsilon u^2\,dx + \frac{1}{2c}\int w_\epsilon u_{xx}^2\,dx.\end{aligned}$$

Picking $c = \epsilon^2$ we find that

$$\frac{d}{dt}\|u(t)\|^2_{L^2_{w,\epsilon}} \le \int w_\epsilon \left(\varphi'(x) + \epsilon \tanh(\epsilon x)\varphi(x) + 3\epsilon^4\right) u^2(x)\,dx.$$

We have shown that a sufficient condition for the derivative to be decreasing at the rate $a\|u\|^2_{L^2_{w,\epsilon}}$ is

$$\Psi(x) \equiv \varphi'(x) + \epsilon \tanh(\epsilon x)\varphi(x) + 3\epsilon^4 < -a < 0.$$

We note that the balancing that occurs here—between $\varphi(x)$, $\varphi'(x)$, and $3\epsilon^4$—is expected. As a rule, exponential weights tend to destabilize parabolic PDEs [2].

Therefore, the $3\epsilon^4$ which comes from the parabolic part of (1.1) *should* tend to destabilize the problem.

The other two terms come from the fundamentally hyperbolic part of the problem. If we were dealing with the equation $u_t+\varphi(x)u_x=0$, then the fact that $\varphi(x)\to\mp 1$ as $x\to\pm\infty$ would force the solutions to "flow in from infinity." Thus, the term involving $\varphi(x)$ *should* help us. Similarly, if we were dealing with the hyperbolic PDE, then the solution would "focus" where $\varphi'(x)$ is negative, and hence, its $L^2_{w_\epsilon}$ norm would shrink at a rate related to $\varphi'(x)$. The solution would "spread" where $\varphi'(x)$ is positive, and there would be an increase in the $L^2_{w_\epsilon}$ norm.

The condition on $\Psi(x)$ says that the "focusing" (or "spreading") due to $\varphi'(x)$, the drift in from infinity due to $\varphi(x)$, and the smearing caused by the parabolic term must balance properly. If they balance properly, then we can show that the solutions of (1.1) decay in our weighted norm. In the next section we will show that $\Psi(x)\le -a$ by using various estimates on the size of $\varphi(x)$ and $\varphi'(x)$ for various regions of the real line. That is, we will show that for certain values of ϵ the three effects balance properly.

Remark. If we show that $\Psi(x)\le -a$, then we will have shown that $\|u\|_{L^2_{w_\epsilon}}\le e^{-at/2}\|u_0\|_{L^2_{w_\epsilon}}$. This is a more precise result than our result in [2]. Using a combination of the theory of sectorial operators and Michelson's results the most that one can say is that there exists a constant C such that $\|u\|_{L^2_{w_\epsilon}}\le Ce^{-at/2}\|u_0\|_{L^2_{w_\epsilon}}$. In this sense, the analytical method gets better results than our previous method got.

3. The proof that $\Psi(x)<-a$. First we note that as $\tanh(\epsilon x)$ and $\varphi(x)$ are odd and $\varphi'(x)$ is even, $\Psi(x)$ is even. Therefore, if $\Psi(x)<-a$ in the region $x\ge 0$, then $\Psi(x)<-a$ on the whole real line.

Looking at the computer drawn plot of $\varphi(x)$ (for purposes of illustration) it is clear that if $\Psi(x)$ is to be negative for small x, then φ' must be rather negative for small x. Because $\varphi(x)$ itself is small for small x, the $\varphi'(x)$ must be quite negative in order to force $\Psi(x)$ to be negative. Similarly, for large x it is clear that $\varphi'(x)$ is small. Therefore, for large x we will have to show that $\varphi(x)$ is relatively large and negative. From Figure 3.1, it seems that it should be possible to prove these results.

There are two major steps in our proof. First, we extend the methods of Koppel and Howard. They proved the existence of a solution $\varphi(x)$ by showing that two curves must intersect. We determine how the curves intersect. We then show that in each of four regions $\varphi(x)$ and $\varphi'(x)$ are small enough that $\Psi(x)$ is negative in those four regions. The four regions are as follows:

1. Region I, which goes from $x=0$ until the first point at which $\varphi(x)=-.5$.
2. Region II, which goes from the first point at which $\varphi(x)=-.5$ to the first point at which $\varphi(x)=-1$.
3. Region III, which goes from the first point at which $\varphi(x)=-1$ to the first point at which $\varphi'(x)=0$.
4. Region IV, which is the region to the right of Region III.

(In Figure 3.1, the four regions are delineated by dotted lines. The first region starts to the right of the first dotted line.) In all of these steps we make use of estimates contained in Koppel and Howard.

3.1. Locating the intersection—An introduction. Before proceeding, we explain the nature of the method used by Koppel and Howard. (Note that their results are stated for $-\varphi(t)$. We convert all of their results to results about $\varphi(x)$.) Following Koppel and Howard, we denote $\frac{d}{dx}$ by $\dot{}$ and $\frac{d^2}{dx^2}$ by $\ddot{}$. Koppel and Howard

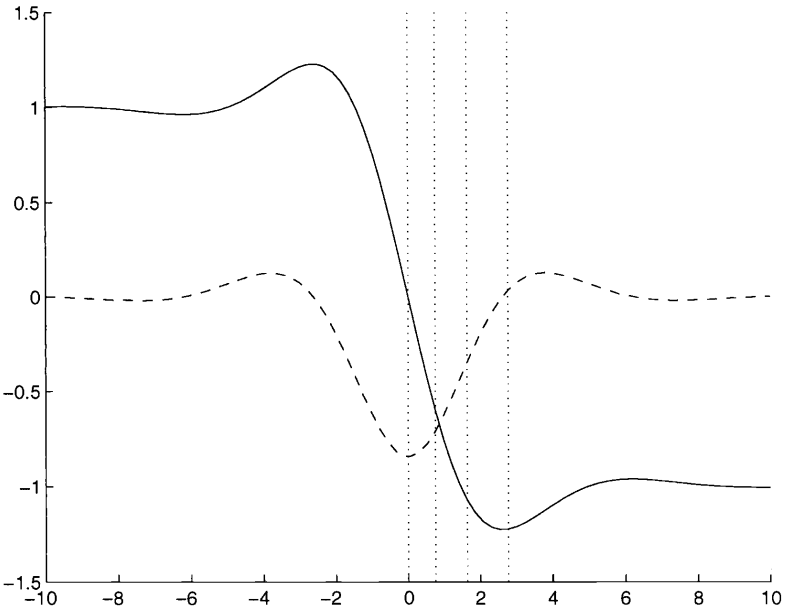


FIG. 3.1. $\varphi(x)$ *(solid line) and* $\varphi'(x)$ *(dashed line).*

look at solutions of the equation

$$\frac{d^3}{dx^3}p(x,\nu) = \frac{1}{2}(1 - p^2(x,\nu)) \tag{3.1}$$

subject to the initial conditions $p(0,\nu) = \ddot{p}(0,\nu) = 0, \dot{p}(0,\nu) = -\nu$. They find analytical bounds for $p(x,\nu)$ and its first two derivatives. They show that the function $p_2(x,\nu)$ is an upper-bound of $p(x,\nu)$ in the region $0 \le x \le 2\sqrt{\nu}$ (see fact I.1 below) and that $Q_3(x,\nu) = p_2(x,\nu) - 6\nu^3 x^9/9!$ is a lower-bound as long as $p(x,\nu)$ has not yet hit -1 (see fact I.2 below).

They proceed to show that the stable manifold, $u(x,\kappa)$ (where κ parameterizes the manifold), of the solution $(-1,0,0)$ can be approximated by $-1 + u_2(x,\kappa)$ where $u_2(x,\kappa)$ is defined below (in fact II.5). Looking at fact II.5, we see that the parameter κ is essentially the "amplitude" of the perturbation. They proceed to estimate the first point at which p hits -1 and the first point after $x = 0$ at which u hits -1. They show that there exists a ν_0 and a κ_* and points x_0 and x_* such that $p(x_0,\nu_0) = (-1,\alpha,\beta)$ and $u(x_*,\kappa_*) = (-1,\alpha,\beta)$. Since the equation is autonomous, this shows that there exists an odd solution of the ODE that is equal to $(0,-\nu,0)$ when $x = 0$ and that tends towards -1 as x gets large.

Koppel and Howard do not try to evaluate ν_0 and κ_*. As we need to make use of particular properties of the solution, we must know these values more precisely.

In our proof we make use of the following facts proven in Koppel and Howard [4].

Facts from Koppel and Howard.

I. Facts related to $p(x,\nu)$.

1. The function $p_2(x,\nu) = -\nu x + x^3/12 - \nu^2 x^5/120 + \nu x^7/2520 - x^9/145,152$ is an upper-bound of p in the region $0 \le x \le 2\sqrt{\nu}$. Furthermore, in this region $\dot{p}_2$ and $\ddot{p}_2$ are upper-bounds of the first two derivatives of p as well [4, p. 318].

2. The function $Q_3(x,\nu) = p_2(x,\nu) - 6\nu^3 x^9/9!$ is a lower-bound for $p(x,\nu)$ as long as $p(x,\nu)$ has not yet hit -1. Furthermore, in this region $\dot{Q}_3$ and $\ddot{Q}_3$ are lower-bounds of the first two derivatives of p as well [4, p. 319].
3. For all $x \geq 0$, $\dot{p}(x,\nu) \leq -\nu + x^2/4$ [4, p. 318, Eq. 6.1].
4. Let $x_{0,\nu}$ be the first point at which $p(x,\nu) = 0$. Let T be the curve of points $(\dot{p}(x_{0,\nu},\nu), \ddot{p}(x_{0,\nu},\nu))$. Then T slopes upward [4, p. 321].

II. Facts related to $u(x,\kappa)$.

5. The function

$$u_2(x,\kappa) = -\kappa e^{-x/2}\sin\left(\frac{\sqrt{3}}{2}x\right) + \frac{\kappa^2}{8}e^{-x}\left(1 + \frac{2}{7}\cos(\sqrt{3}x)\right)$$

satisfies the inequality

$$|u(x,\kappa) - (-1 + u_2(x,\kappa))| \leq .0714\kappa^3 e^{-3x/2} \quad [4, \text{p. } 327].$$

6. Similarly, the function

$$\dot{u}_2(x,\kappa) = -\kappa e^{-x/2}\cos\left(\frac{\sqrt{3}}{2}x + \frac{\pi}{6}\right) - \frac{\kappa^2}{8}e^{-x}\left(1 + \frac{4}{7}\cos\left(\sqrt{3}x - \frac{\pi}{3}\right)\right)$$

satisfies

$$|\dot{u}(x,\kappa) - \dot{u}_2(x,\kappa)| \leq .1363\kappa^3 e^{-3x/2} \quad [4, \text{p. } 327\text{–}328].$$

7. Let $x_{0,\kappa}$ be the first nonnegative point at which $u(x,\kappa) = 0$. Let S_1 be the curve of points $(\dot{u}(x_{0,\kappa},\kappa), \ddot{u}(x_{0,\kappa},\kappa))$. Then for $0 \leq \kappa \leq .856$, S_1 slopes downward [4, p. 330].
8. For $0 \leq \kappa \leq 1$ the curve S_1 lies between the curves $y = -x$ and $y = -x + .175x^2 - .3x^3$ [4, p. 329].
9. For any $\kappa \leq 1$, $u(x,\kappa)$ crosses zero exactly once in the region $0 \leq x \leq .3\kappa$ [4, Lem. 6.5, p. 328].

3.2. A refinement of the method of Koppel and Howard. In their paper, Koppel and Howard establish the existence of $\varphi(x)$ by showing that the manifold of the solutions of (3.1) that have initial data $p(0,\nu) = \ddot{p}(0,\nu) = 0, \dot{p}(0,\nu) = -\nu$ intersects the stable manifold of the solution $u \equiv -1$. We are interested in actually figuring out roughly how the intersection occurs.

From fact II.8, we see that if we can estimate $\dot{u}(x_{0,\nu},\nu)$, then we automatically have an estimate of $\ddot{u}(x_{0,\nu},\nu)$ as well. From facts II.7 and II.8 we see that if we can locate two points on S_1, we can say quite a bit about the locations of all the points on S_1 between these two points. We note that if $\kappa = 0$ then $u_2 \equiv 0$. Thus, the intersection with -1 occurs when $\dot{u} = \ddot{u} = 0$. Therefore, $(0,0)$ is one point on S_1.

Next, we would like to estimate the point on S_1 that corresponds to $\nu = .67$. Using fact II.5, it is easy to see that $u(.09,.67) > -1$ and $u(.17,.67) < -1$. Thus, $u(x,.67)$ must cross -1 somewhere in the region $.09 \leq x \leq .17$. Making use of fact II.6, we see that the function $\dot{u}_2 + .1363\kappa^3 e^{-3x/2}$ is an upper bound for $\dot{u}(x,\kappa)$. It is easy to see that $\dot{u}_2 + .1363\kappa^3 e^{-3x/2}$ is an increasing function of x if $\kappa \leq .7$ and $x \leq .2$. (Just check its derivative.) Therefore, we evaluate $\dot{u}_2 + .1363\kappa^3 e^{-3x/2}$ at $x = .17$ in order to get an upper-bound on the derivative at the point at which $u(x,.67)$ first

crosses -1. We find that $\dot{u}(x, .67) \leq -.5174$ as long as $x \leq .17$. Using facts II.7 and II.8 we see that any curve that must cross the crosshatched region in Figure 3.2 must cross the curve S_1, and $0 \leq \kappa \leq .67$ at the point at which the curves cross.

We now show that T crosses the crosshatched region for a value of ν between .8 and .9. If $\nu = .8$, then at the point at which p crosses -1, Koppel and Howard have shown that $-.2831 \leq \dot{p} \leq -.2770$ [4, p. 319–320] and $.472 \leq \ddot{p} \leq .493$. Following Koppel and Howard we find a similar estimate for $\nu = .9$. We note that $Q_3(1.28, .9) = -.9986$ and $p_2(1.29, .9) = -1.0042$. As Q_3 is a lower-bound, and p_2 is an upper-bound, we see that there exists $1.28 \leq x_{0,.9} \leq 1.29$ such that $p(x_{0,.9}, .9) = -1$. A simple calculation shows that if $\nu = .9$, then both $\ddot{p}_2$ and $\ddot{Q}_3$ are positive in the $1.28 \leq x \leq 1.29$. Therefore, both $\dot{p}_2$ and $\dot{Q}_3$ are increasing in this region. Thus,

$$-.5712 = \dot{Q}_3(1.28, .9) \leq \dot{p}(x_{0,.9}, .9) \leq \dot{p}_2(1.29) = -.5664.$$

As p is decreasing when it passes through $x_{0,.9}$, we see from (3.1) that the third derivative of p goes from positive to negative as it crosses through $x_{0,.9}$. That is, $x_{0,.9}$ is a local maximum of $\ddot{p}(x, .9)$. Thus, $\ddot{p}(x_{0,.9}, .9) \geq \ddot{Q}_3(1.28, .9) = .4007$. Note that $\frac{d^3}{dx^3} p_2 = \frac{1}{2}(1 - p_1^2)$ with $p_1(x, \nu) = -\nu x + x^3/12$. Clearly, p_1 is decreasing in the interval, and we see that $p_1(1.29, .9) = -.9821 > -1$. Thus, we find that the third derivative of p_2 is positive in the whole interval. That implies that $\ddot{p}_2$ is increasing in the interval. Thus, $\ddot{p}(x_{0,.9}, \nu) \leq \ddot{p}_2(1.29, .9) = .4058$. We see then that $.4007 \leq \ddot{p} \leq .4058$.

Remark. In evaluating p_2 and Q_3, we are evaluating polynomials with rational coefficients at rational numbers. Such calculations can be made with perfect accuracy. A simple way to keep such calculations exact is to do the calculations using Mathematica. All critical calculations involving polynomials with rational coefficients were performed in this fashion. Only after the calculations were finished were the results rounded to four places.

Looking at Figure 3.2 we see that because T slopes up (see fact I.4), T must cross the crosshatched region.

Thus, we see that the intersection of T and S_1 takes place when κ is between 0 and .67 and ν is between .8 and .9.

The behavior of $\varphi(x)$ can now be studied by studying the behavior of $p(x, \nu)$ *until* its first positive -1 crossing and by studying the behavior of $u(x, \kappa)$ *after* its first -1 crossing. In section 3.3 we study the behavior of $\varphi(x)$ in regions I and II by making use of our knowledge of $p(x, \nu)$. In section 3.4 we study the behavior of $\varphi(x)$ by making use of our knowledge of $u(x, \kappa)$.

3.3. Estimates for regions I and II. From fact I.3, we see that for any ν we know that $p_2(x, \nu)$ decreases until at least $x = 2\sqrt{\nu}$. We want to find a value, $x_\dagger$, such that for all $.8 \leq \nu \leq .9$ we know $p_2(x_\dagger, \nu)$ is less than $-.5$. We will then estimate the values of the derivative in the region $0 \leq x \leq x_\dagger$—a region that includes region I. By showing that in region I the derivative is suitably small, we will later be able to show that $\Psi(x)$ is negative.

If $.8 \leq \nu \leq .9$, we find that

$$\begin{aligned} p_2 &= -\nu x + x^3/12 - v^2 x^5/120 + \nu x^7/2520 - x^9/145,152 \\ &\leq -.8x + x^3/12 - .64x^5/120 + .9x^7/2520 - x^9/145,152. \end{aligned}$$

From fact I.3, we know that $p(x, \nu)$ decreases as long as $x \leq 2\sqrt{\nu} \leq 2\sqrt{.8} = 1.79$. We find that the right-hand side first falls below $-.5$ for some $x \leq .66$; we let $x_\dagger = .66$. From fact I.3, we see that for all $0 \leq x \leq x_\dagger$, the derivative of the solution is less

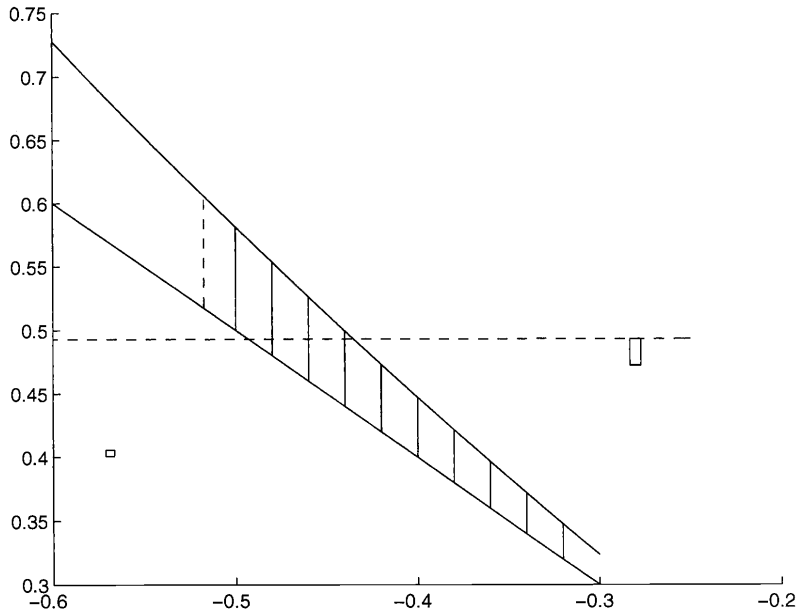


FIG. 3.2. *The crossing of T and S_1. The small box to the left shows the rigorous bounds on $(\dot{p}(x_{0,.8},.8),\ddot{p}(x_{0,.8},.8))$. The small box on the right shows the rigorous bounds on $(\dot{p}(x_{0,.9},.9),\ddot{p}(x_{0,.9},.9))$. The crosshatched region must contain the curve S_1, and T must start from the box on the left and move up to the box on the right. Clearly, T crosses the crosshatched region and, therefore, S_1.*

than or equal to $-\nu + .66^2/4 \le -.6911$. Also, we find that the right-hand side is less than -1 when $x = 1.66$. Therefore, the first -1 crossing occurs before this point. For $0 \le x \le 1.66$ and $.8 \le \nu \le .9$, we find that $p'(x,\nu) \le -.8 + 1.66^2/4 = -.1111$.

Because of the fact that $\tanh(\epsilon x)$ appears in $\Psi(x)$, it will be important to know what the earliest point at which the profile may hit $-.5$ or -1 is. To determine this, we look for the last point at which $Q_3(x)$ cannot be less than $-.5$ or -1 for values of ν between .8 and .9.

$$\begin{aligned} Q_3 &= -\nu x + x^3/12 - \nu^2 x^5/120 + \nu x^7/2520 - x^9/145,152 - 6\nu^3 x^9/9! \\ &\ge -.9x + x^3/12 - .81x^5/120 + .8x^7/2520 - x^9/145,152 - 6.7290x^9/9!. \end{aligned}$$

We find that when $x = .55$ the right-hand side is equal to $-.4815$. Thus, the first $-.5$ crossing does not come before $x = .55$. When $x = 1.2$, the right-hand side is $-.9516$. Thus, the first -1 crossing cannot come before $x = 1.2$.

3.4. Estimates for region III and region IV. We show that as long as $\kappa \le .67$, $\varphi'(x)$ cannot get too positive in region IV. (By definition, $\varphi'(x)$ is nonpositive in region III.) From fact II.6 we see that

$$\begin{aligned} \dot{u} &\le -\kappa e^{-x/2}\cos\left(\frac{\sqrt{3}}{2}x + \frac{\pi}{6}\right) - \frac{\kappa^2}{8}e^{-x}\left(1 + \frac{4}{7}\cos\left(\sqrt{3}x - \frac{\pi}{3}\right)\right) + .1363\kappa^3 e^{-3x/2} \\ &\le -\kappa e^{-x/2}\cos\left(\frac{\sqrt{3}}{2}x + \frac{\pi}{6}\right) + .67^2\left(-\frac{1}{8} + \frac{1}{14} + .1363 \times .67\right)e^{-x} \end{aligned}$$

Region	Does not begin before $x =$	$\varphi \leq$	$\varphi' \leq$
I	0	0	$-.6911$
II	.55	$-.5$	$-.1111$
III	1.2	$-.8292$	0
IV	2.208	$-.8979$	.1783

TABLE 3.1
A summary of our results about φ and φ'.

$$\leq -\kappa e^{-x/2}\cos\left(\frac{\sqrt{3}}{2}x+\frac{\pi}{6}\right)+.0169e^{-x}.$$

A simple calculus argument shows that

$$-e^{-x/2}\cos\left(\frac{\sqrt{3}}{2}x+\frac{\pi}{6}\right)\leq -e^{-\frac{2\pi}{3\sqrt{3}}}\cos(5\pi/6)=.1732.$$

As $-\kappa^2 e^{-x}/8+.1363\kappa^3 e^{-3x/2}$ is always negative if $x\geq 0$ and $\kappa\leq .67$, it is clear that for $x\geq 0$ $\dot{u}$ cannot get positive until one of the sinusoids above gets negative. It is easy to see that the first point at which this happens is $x=\frac{2\pi}{3\sqrt{3}}\geq 1.209$. We will need this fact later as well. We find that the largest value of $\dot{u}$ cannot exceed $.1732+.0169e^{-1.209}=.1783$.

Looking at u region III and making use of fact II.5, we find that

$$\begin{aligned}u&\leq -1-\kappa e^{-x/2}\sin\left(\frac{\sqrt{3}}{2}x\right)+\frac{\kappa^2}{8}e^{-x}\left(1+\frac{2}{7}\cos(\sqrt{3}x)\right)+.0714\kappa^3e^{-3x/2}\\&\leq -1-\kappa e^{-x/2}\sin\left(\frac{\sqrt{3}}{2}x\right)+e^{-x}\frac{9\kappa^2}{56}+.0714\kappa^3e^{-3x/2}\\&\leq -1-\kappa e^{-x/2}\sin\left(\frac{\sqrt{3}}{2}x\right)+.0963.\end{aligned}$$

A simple calculus argument shows that $-\kappa e^{-x/2}\sin(\frac{\sqrt{3}}{2}x)\leq .0771$. This shows that $u\leq -1+.1708=-.8292$. Thus, we see that in region III the value of φ never gets above $-.8292$. The region in which φ' may be positive, region IV, starts no earlier than when the x used in u_2 is equal to 1.209. In this region we find that $\varphi(x)$ never gets above $-1+.0771+e^{-1.209}\frac{9}{56}.67^2+.0714(.67^3)e^{-3\cdot 1.209/2}=-.8979$ and φ' never gets above .1783.

From fact II.9 we see that the first -1 crossing of u never occurs later than $.3\cdot .67=.201$. We have seen that the derivative never gets positive earlier than 1.209. Thus, the derivative never gets positive earlier than 1.008 units past the point at which the solution first crosses -1. Combining this with our earlier result, we find that the $\varphi'(x)$ never gets positive before $x=2.208$. We summarize these results in Table 1.

We now evaluate $\Psi(x)$ in each of our regions. Recall that $\Psi(x)=\varphi'(x)+\epsilon\tanh(\epsilon x)\varphi(x)+3\epsilon^4$. In region I, we see that $\Psi(x)\leq -.6911+3\epsilon^4$. In region II,

we know that $x \geq .55$. Thus, in region II $\Psi(x) \leq -.1111 - .5\epsilon \tanh(.55\epsilon) + 3\epsilon^4$. In region III, $\Psi(x) \leq -.8292\epsilon \tanh(1.2\epsilon) + 3\epsilon^4$. Finally, in region IV we find that $\Psi(x) \leq .1783 - .8979\epsilon \tanh(2.2\epsilon) + 3\epsilon^4$.

We find that $-.6911 + 3\epsilon^4$ is negative as long as $\epsilon \leq .69$. Similarly, $-.1111 - .5\epsilon \tanh(.55\epsilon) + 3\epsilon^4$ is negative as long as $\epsilon \leq .49$. Additionally, $-.8292\epsilon \tanh(1.2\epsilon) + 3\epsilon^4$ is negative until $\epsilon > .54$. Finally, $.1783 - .8979\epsilon \tanh(2.2\epsilon) + 3\epsilon^4$ is negative as long as $.43 \leq \epsilon \leq .47$. Thus, if $.43 \leq \epsilon \leq .47$ we find that Ψ is negative and all solutions of the PDE must decay exponentially quickly in the w_ϵ norm.

We have shown that there exist values of ϵ for which solution of (1.1) decay exponentially quickly in the w_ϵ norm. Using Michelson's rigorous numerical results, one can show that a larger range of ϵ gives this decay. That these methods give somewhat better results is not surprising—the estimates we use here are rather crude. Conversely, we have seen that where our method works, it gives us somewhat more information than we got using the numerical results. Using our current method, we find that there is no possibility of the solution increasing even momentarily. Our results also have the advantage of all analytical results. They are easy to understand and simple to check.

Acknowledgment. We would like to thank Professor Steve Schochet of Tel Aviv University for many helpful discussions and for suggesting that it was possible to show that there exist epsilon for which $\Psi(x)$ is strictly negative.

REFERENCES

[1] J. Alexander, R. Gardner, and C. K. R. T. Jones, *A topological invariant arising in the stability analysis of traveling waves*, J. Riene Angew. Math., 410 (1990), pp. 167–212.

[2] S. Engelberg, *The stability of the viscous shock profiles of the Burgers' equation with a fourth order viscosity*, Comm. Partial Differential Equations, 21 (1996), pp. 889–922.

[3] J. Goodman, *Nonlinear asymptotic stability of viscous shock profiles for conservation laws*, Arch. Rational Mech. Anal., 95 (1986), pp. 325–344.

[4] N. Kopell and L. N. Howard, *Bifurcations and Trajectories Joining Critical Points*, Adv. Math., 18 (1975), pp. 306–358.

[5] T. P. Liu, *Nonlinear stability of shock waves for viscous conservation laws*, Bull. Amer. Math. Soc. (N.S.), 12 (1985), pp. 233–236.

[6] 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.

[7] C. K. McKord, *Uniqueness of Connecting Orbits in the Equation* $y^{(3)} = y^2 - 1$, Math. Anal. Appl., 114 (1986), pp. 584–592.

[8] D. Michelson, *Stability of Bunsen flame profiles in the Kuramoto–Sivashinsky equation*, SIAM J. Math. Anal. 27 (1996), pp. 765–781.

[9] M. S. Mock, *On fourth-order dissipation and single conservation laws*, Comm. Pure Appl. Math., 29 (1976), pp. 383–388.

[10] G. Sivashinsky, *Nonlinear analysis of hydrodynamic instability in laminar flames*, Acta Astronautica, 4 (1977), pp. 1117–1206.

[11] A. Szepessy and Z. Xin, *Nonlinear Stability of Viscous Shock Waves*, Arch. Rational Mech. Anal., 122 (1993), pp. 53–103.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 937–948

BEHAVIOR OF SOLUTIONS OF 2D QUASI-GEOSTROPHIC EQUATIONS*

PETER CONSTANTIN† AND JIAHONG WU‡

Abstract. We study solutions to the 2D quasi-geostrophic (QGS) equation

$$\frac{\partial\theta}{\partial t} + u\cdot\nabla\theta + \kappa(-\Delta)^{\alpha}\theta = f$$

and prove global existence and uniqueness of smooth solutions if $\alpha \in (\frac{1}{2}, 1]$; weak solutions also exist globally but are proven to be unique only in the class of strong solutions. Detailed aspects of large time approximation by the linear QGS equation are obtained.

Key words. quasi-geostrophic equation, existence, uniqueness, large time approximation

AMS subject classifications. 76U05, 35Q35

PII. S0036141098337333

1. Introduction.

This paper is concerned with the 2D surface quasi-geostrophic (QGS) equation

$$\frac{\partial\theta}{\partial t} + u\cdot\nabla\theta + \kappa(-\Delta)^{\alpha}\theta = f,$$

where $\alpha \in [0,1]$, $\kappa > 0$, and $\theta = \theta(x,t)$ is a real scalar function of two space variables x and a time variable t. The velocity $u = (u_1, u_2)$ is incompressible and determined from θ by a stream function ψ:

$$(u_1, u_2) = \left(-\frac{\partial\psi}{\partial x_2}, \frac{\partial\psi}{\partial x_1}\right), \tag{1.1}$$

and the stream function ψ satisfies

$$(-\Delta)^{\frac{1}{2}}\psi = -\theta. \tag{1.2}$$

The nonlocal operator $(-\Delta)^{\beta}$ $(\beta \geq 0)$ is defined through the Fourier transform

$$\widehat{(-\Delta)^{\beta} f}(\xi) = |\xi|^{2\beta}\widehat{f}(\xi),$$

where $\widehat{f}$ is the Fourier transform of f [11]. For notational convenience, we write Λ for $(-\Delta)^{\frac{1}{2}}$.

The variable θ in the 2D QGS equation represents the potential temperature, u is the fluid velocity, and the stream function ψ can be identified with the pressure. When the fractional power $\alpha = 1/2$, the equation, derived from the more general quasi-geostrophic models (see pages 345–368 and 653–670 of [7]), describes the evolution

*Received by the editors April 13, 1998; accepted for publication (in revised form) November 18, 1998; published electronically June 29, 1999.

http://www.siam.org/journals/sima/30-5/33733.html

†Department of Mathematics, The University of Chicago, Chicago, IL 60637 (const@cs.uchicago.edu). The work of this author was partially supported by NSF/DMS grant 9207080.

‡School of Mathematics, Institute for Advanced Study, Princeton, NJ 08540. Current address: Department of Mathematics, University of Texas, Austin, TX 78712-1082 (jiahong@math.utexas.edu). The work of this author while at the Institute for Advanced Study was partially supported by NSF/DMS grant 9304580.

of the temperature on the 2D boundary of a rapidly rotating half-space with small Rossby and Ekman numbers. Dimensionally, the 2D QGS equation with $\alpha = 1/2$ is the analogue of the 3D Navier–Stokes equations. The general fractional power α is considered here in order to observe the minimal power of Laplacian necessary in the analysis and thus make a comparison with the 3D Navier–Stokes equations [3], [6].

Recently, this equation has been intensively investigated because of both its mathematical importance and its potential for applications in meteorology and oceanography [4], [7], [5]. Mathematically, the behavior of solutions to the 2D QGS equation is strikingly similar to that of the potentially singular solutions to the 3D hydrodynamics equations. Despite exhibiting a number of similar features, the 2D QGS equation is considerably simpler than the 3D Euler or Navier–Stokes equations.

The smooth solution of the QGS equation is unique but, if $\kappa = 0$, it is known to exist only for a finite time [4]. On the other hand, weak solutions are global but their uniqueness is unknown [8]. Whether the smooth solution develops singularity in finite time and whether weak solutions are unique are fundamental mathematical issues related to the QGS equation. We show in section 2 that the solution remains smooth for all time for $\alpha \in (\frac{1}{2}, 1]$ and any weak solution must coincide with a more regular solution as long as such a strong solution exists.

Large time behavior of weak solutions is investigated in sections 3 and 4. In section 3, the L^2 decay rate of order $t^{-\frac{1}{2\alpha}}$ is obtained. For generic initial data, this rate is optimal. The solution θ of the nonlinear equation may be approximated by the solution Θ of the linear equation with a higher-order correction. An explicit form for the higher-order correction is attempted in section 4. A rate of order $t^{\frac{1}{2}-\frac{1}{\alpha}}$ is first obtained without any smoothness assumption. With the assumption that

$$\|\Lambda^{2-2\alpha+\delta}\theta(\cdot,t)\|_{L^2} \quad \sim \quad t^{-\epsilon},$$

the ratio and the difference are shown to behave as follows:

$$\frac{\|\theta(\cdot,t)\|_{L^2}}{\|\Theta(\cdot,t)\|_{L^2}} \sim 1 + O(t^{-\min\{\frac{1}{2\alpha},\epsilon\}}), \quad \|\theta(\cdot,t) - \Theta(\cdot,t)\|_{L^2} \sim t^{-\frac{1}{2\alpha}-\min\{\frac{1}{2\alpha},\epsilon\}},$$

which imply that the effect of the nonlinearity is felt only in the higher-order correction.

We conclude this introduction by mentioning the global existence result of weak solutions obtained in [8]. When not specified, the spatial domain can be either the whole $\mathbb{R}^2$ or the 2D torus $\mathbb{T}^2$.

PROPOSITION 1.1. *Let $T > 0$ be arbitrary. Then for every $\theta_0 \in L^2$ and $f \in L^2([0,T]; H^{-\alpha})$, there exists a weak solution of*

$$\partial_t\theta + u \cdot \nabla\theta + \kappa\Lambda^{2\alpha}\theta = f, \tag{1.3}$$

$$\theta\,|_{t=0} = \theta_0 \tag{1.4}$$

which satisfies

$$\theta \in L^\infty([0,T]; L^2) \cap L^2([0,T]; H^\alpha).$$

2. Global smooth solution and uniqueness. It is shown here that weak solutions of the QGS equation are globally smooth for $\alpha \in (\frac{1}{2}, 1]$ and "strong" solutions are unique. The spatial domain here is the 2D torus $\mathbb{T}^2$.

THEOREM 2.1. *Let $\alpha \in (\frac{1}{2}, 1]$, $\beta \geq 0$, and $\beta + 2\alpha > 2$. If $\theta_0 \in H^\beta(\mathbb{T}^2)$ and if, for $T > 0$,*

$$f \in L^2([0,T]; H^{\beta-\alpha}), \quad \int_0^T \|f(\tau)\|_{L^q} d\tau < \infty,$$

where $q = \infty$ for $\beta \geq 1$ and $q = 2/(1-\beta)$ for $\beta < 1$, then the solution θ of (1.3) and (1.4) obeys for all $t \leq T$

$$\|\Lambda^\beta \theta(t)\|_{L^2} \leq C, \tag{2.1}$$

where C is constant depending only on T, $\|\theta_0\|_{H^\beta}$, $\|f\|_{L^2([0,T];H^{\beta-\alpha})}$, and $\int_0^T \|f(\tau)\|_{L^q} d\tau$.

Proof. We sketch the proof. Taking the scalar product of (1.3) with $\Lambda^{2\beta}\theta$

$$\frac{1}{2}\frac{d}{dt}\int |\Lambda^\beta \theta|^2 + \kappa \int |\Lambda^{\alpha+\beta}\theta|^2 = -\int (u \cdot \nabla\theta)\Lambda^{2\beta}\theta + \int \Lambda^{2\beta}\theta f$$

and using the estimates

$$\left| \int \Lambda^{2\beta}\theta f \right| \leq \frac{\kappa}{4}\|\Lambda^{\alpha+\beta}\theta\|_{L^2}^2 + \frac{1}{\kappa}\|\Lambda^{\beta-\alpha} f\|_{L^2}^2, \tag{2.2}$$

$$\left| \int (u \cdot \nabla\theta)\Lambda^{2\beta}\theta \right| \leq \frac{\kappa}{4}\|\theta\|_{H^{\alpha+\beta}}^2 + C(\kappa, \theta_0, f)\|\theta\|_{H^\beta}^2, \tag{2.3}$$

where $C(\kappa, \theta_0, f)$ is constant, we obtain (2.1) after applying Gronwall's inequality. The estimate (2.3) is obtained by using the calculus inequality (see page 61 of [8] and inequality (3.1.59) on page 74 of [12])

$$\|\Lambda^s(gh)\|_{L^2} \leq C(\|g\|_{L^q}\|\Lambda^s h\|_{L^p} + \|h\|_{L^q}\|\Lambda^s g\|_{L^p})$$

with $1/p + 1/q = 1/2$, $g = u$, $h = \theta$, $s = \beta + 1 - \alpha$, and the maximum principle

$$\|\theta\|_{L^q} \leq \|\theta_0\|_{L^q} + \int_0^t \|f(\tau)\|_{L^q} d\tau. \qquad \square$$

Although weak solutions may not be unique, there is at most one solution in the class of "strong" solutions.

THEOREM 2.2. *Assume that $\alpha \in (\frac{1}{2}, 1]$, $T > 0$, p and q satisfy*

$$p \geq 1, \quad q > 0, \qquad \frac{1}{p} + \frac{\alpha}{q} = \alpha - \frac{1}{2}; \tag{2.4}$$

then there is at most one solution θ of the QGS equation with initial data $\theta_0 \in L^2$ such that

$$\theta \in L^\infty([0,T]; L^2) \cap L^2([0,T]; H^\alpha), \tag{2.5}$$

$$\theta \in L^q([0,T]; L^p). \tag{2.6}$$

We make two remarks.

Remark 2.3. It is clear from the proof given below that we can assume that only one of the two solutions is "strong," i.e., in the class (2.5), (2.6), the other being only a weak solution.

Remark 2.4. By taking $\alpha = 1$, (2.6) with (2.4) reduces exactly to the regularity assumptions in obtaining uniqueness for weak solutions to the 3D Navier–Stokes equations (cf. Temam [13, p. 299]). Theorem 2.2 is a sort of generalization in the sense that it holds for a range of $\alpha \in (\frac{1}{2}, 1]$.

Proof of Theorem 2.2. The difference $\theta = \theta_A - \theta_B$ of two solutions θ_A and θ_B satisfies

$$\partial_t \theta + u \cdot \nabla \theta_A + u_B \cdot \nabla \theta + \kappa \Lambda^{2\alpha} \theta = 0 \tag{2.7}$$

in which $u = u_A - u_B$ with u_A and u_B being the velocities corresponding to θ_A and θ_B. We take the scalar product of (2.7) with $\psi = -\Lambda^{-1}\theta$ and use

$$\int_{\mathbb{T}^2} \psi u \cdot \nabla \theta_A = 0,$$

$$\left| \int_{\mathbb{T}^2} \theta u_B \cdot \nabla \psi \right| \le \kappa \|\psi\|^2_{H^{\alpha+\frac{1}{2}}} + C(\kappa) \|\theta_B\|^{\frac{1}{1-\beta}}_{L^p} \|\psi\|^2_{H^{\frac{1}{2}}},$$

where $\beta = \frac{1}{\alpha}\left(\frac{1}{2} + \frac{1}{p}\right)$ and $C(\kappa) = C\kappa^{-\frac{\beta}{1-\beta}}$ (see page 32 of [8]). It then follows that

$$\frac{d}{dt}\|\psi\|^2_{H^{\frac{1}{2}}} \le C(\kappa) \|\theta_B\|^{\frac{1}{1-\beta}}_{L^p} \|\psi\|^2_{H^{\frac{1}{2}}},$$

which implies that $\psi = 0$ and thus $\theta = 0$. □

3. Large time behavior. The large time behavior of weak solutions is investigated in this section. We adapt well-known ideas of Amick, Bona, and Schonbek [1] and Schonbek [9], [10].

We first analyze the case when the force $f = 0$ and the result can be stated as follows.

THEOREM 3.1. *Let $\alpha \in (0, 1]$ and $\theta_0 \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$. Then there exists a weak solution θ of the 2D QGS equation*

$$\partial_t \theta + u \cdot \nabla \theta + \Lambda^{2\alpha} \theta = 0, \qquad \theta\,|_{t=0} = \theta_0 \tag{3.1}$$

such that

$$\|\theta(\cdot, t)\|_{L^2(\mathbb{R}^2)} \le C(1+t)^{-\frac{1}{2\alpha}}, \tag{3.2}$$

where C is a constant depending on L^1 and L^2 norms of θ_0.

Remark 3.2. For generic initial data, the rate obtained in Theorem 3.1 is optimal, as implied by Theorem 4.6 of section 4.

The proof of Theorem 3.1 consists of two major steps. The first step is a formal argument to show that (3.2) holds for smooth solutions. In the second step the formal argument is applied to a sequence of "retarded mollifications" [2] and we obtain Theorem 3.1 after passing to the limit. We will need a simple estimate.

LEMMA 3.3. *Assume that $\theta_0 \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$. Then θ satisfies the a priori estimate*

$$|\widehat{\theta}(\xi,t)| \le \|\theta_0\|_{L^1} + |\xi| \int_0^t \|\theta(\tau)\|_{L^2}^2 d\tau.$$

Proof. We have from (3.1)

$$(*) \qquad \partial_t \widehat{\theta} + |\xi|^{2\alpha} \widehat{\theta} = -\widehat{u \cdot \nabla \theta}.$$

Since $\nabla \cdot u = 0$,

$$|\widehat{u \cdot \nabla \theta}| \le |\xi| \|\theta(t)\|_{L^2}^2.$$

After integrating $(*)$, we obtain

$$|\widehat{\theta}(\xi,t)| \le |\widehat{\theta_0}(\xi)| + |\xi| \int_0^t \|\theta(\tau)\|_{L^2}^2 d\tau \le \|\theta_0\|_{L^1} + |\xi| \|\theta_0\|_{L^2}^2 t. \qquad \square$$

Proof of Theorem 3.1. Taking the scalar product of (3.1) with θ we obtain

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}^n} |\theta|^2 + \int_{\mathbb{R}^n} (\Lambda^\alpha \theta)^2 = 0.$$

Using Plancherel's theorem,

$$\frac{d}{dt} \int_{\mathbb{R}^2} |\widehat{\theta}|^2 + 2 \int_{\mathbb{R}^2} |\xi|^{2\alpha} |\widehat{\theta}|^2 = 0.$$

For the second term

$$\int_{\mathbb{R}^2} |\xi|^{2\alpha} |\widehat{\theta}|^2 \ge \int_{B(t)^c} |\xi|^{2\alpha} |\widehat{\theta}|^2 \ge g^{2\alpha}(t) \int_{B(t)^c} |\widehat{\theta}|^2$$

$$= g^{2\alpha}(t) \int_{\mathbb{R}^2} |\widehat{\theta}|^2 - g^{2\alpha}(t) \int_{B(t)} |\widehat{\theta}|^2,$$

where $g \in C([0,\infty); \mathbb{R}^+)$ remains to be determined and $B(t)^c$ is the complement of $B(t)$ with

$$B(t) = \left\{ \xi \in \mathbb{R}^2 : \quad |\xi| < g(t) \right\}.$$

By Lemma 3.3, we obtain

$$\frac{d}{dt} \int_{\mathbb{R}^2} |\widehat{\theta}|^2 + 2 g^{2\alpha}(t) \int_{\mathbb{R}^2} |\widehat{\theta}|^2$$

$$(3.3) \qquad \le 2\pi g^{2\alpha}(t) \int_0^{g(t)} \left[\|\theta_0\|_{L^1} + r \int_0^t \|\theta(\tau)\|_{L^2}^2 d\tau \right]^2 r dr.$$

By integrating (3.3), we have

$$e^{2\int_0^t g^{2\alpha}(\tau)d\tau}\int_{\mathbb{R}^2}|\widehat{\theta}|^2 \le \|\theta_0\|_{L^2}^2$$

$$+\int_0^t e^{2\int_0^s g^{2\alpha}(\tau)d\tau}\left[C_1 g^{2\alpha+2}(s)+C_2 s g^{2\alpha+4}(s)\int_0^s \|\theta(\tau)\|_{L^2}^4 d\tau\right]ds, \tag{3.4}$$

where $C_1 = 2\pi\|\theta_0\|_{L^1}^2$ and $C_2 = \pi$.

To obtain a basic estimate, we take $g^{2\alpha}(t) = \left(\frac{1}{2}+\frac{1}{2\alpha}\right)[(e+t)\ln(e+t)]^{-1}$ and thus $e^{2\int_0^t g^{2\alpha}(\tau)d\tau} = [\ln(e+t)]^{(1+\frac{1}{\alpha})}$. We then obtain from (3.4)

$$\|\theta\|_{L^2}^2 \le C[\ln(e+t)]^{-1-\frac{1}{\alpha}}.$$

To obtain the sharp decay result, we take $g^{2\alpha}(t) = \frac{1}{2\alpha(t+1)}$ and thus $e^{2\int_0^t g^{2\alpha}(\tau)d\tau} = (1+t)^{\frac{1}{\alpha}}$. From (3.4),

$$\|\theta(t)\|_{L^2}^2 \le C(t+1)^{-\frac{1}{\alpha}} + C(t+1)^{(1-\frac{2}{\alpha})}\int_0^t \|\theta(s)\|_{L^2}^2[\ln(e+s)]^{-1-\frac{1}{\alpha}}ds.$$

Using Gronwall's inequality and the fact that $\alpha \le 1$,

$$\|\theta(t)\|_{L^2}^2 \le C(1+t)^{-\frac{1}{\alpha}}, \tag{3.5}$$

where the constant C depends on the L^1 and L^2 norms of θ_0. We note here that (3.5) is actually obtained by first taking $g^{2\alpha}(t) = \left(\frac{1}{2\alpha}-\epsilon\right)\frac{1}{1+t}$ and then passing to the limit as $\epsilon \to 0$. This completes the formal argument step.

Next we construct a sequence of retarded mollifications θ_n and carry over the formal arguments to θ_n. We will present here only the main ideas. We approximate the QGS equation by a sequence of equations

$$\partial_t\theta_n + u_n\cdot\nabla\theta_n + \Lambda^{2\alpha}\theta_n = 0, \tag{3.6}$$

where $\delta_n \to 0$ and $u_n = S_{\delta_n}(\theta_n)$ is obtained from θ_n by

$$S_{\delta_n}(\theta_n) = \int_0^\infty \phi(\tau)\mathcal{R}^\perp\theta_n(t-\delta_n\tau)d\tau.$$

We denote here $\mathcal{R}^\perp = (-\partial_{x_2}\Lambda, \partial_{x_1}\Lambda)$ as the Riesz transform. The smooth function ϕ is nonnegative with compact support in $[1,2]$ and $\int_0^\infty \phi(t)dt = 1$. For each n, (3.6) is a linear equation since the values of $u_n(t)$ depend only on the values of θ_n in $[t-2\delta_n, t-\delta_n]$.

Without giving details, we point out that θ_n converges to a weak solution θ strongly in L^2 for almost every t. Hence

$$\|\theta(t)\|_{L^2} \le \|\theta_n(t)-\theta(t)\|_{L^2} + \|\theta_n(t)\|_{L^2} \le C(1+t)^{-\frac{1}{2\alpha}},$$

where C is a constant depending only on the L^1 and L^2 norms of θ_0. This completes the proof of Theorem 3.1. □

We now consider the case when the force f is not zero.

THEOREM 3.4. *Let $\alpha \in (0,1]$ and $\theta_0 \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$. Assume that $f \in L^1([0,\infty); L^2)$, satisfying*

$$\|f(\cdot,t)\|_{L^2} \le C(1+t)^{-\frac{1}{\alpha}-1}, \qquad |\widehat{f(\xi,t)}| \le C|\xi|^\alpha \tag{3.7}$$

for some constant C. Then there is a weak solution of the QGS equation

$$\partial_t\theta + u\cdot\theta + \Lambda^{2\alpha}\theta = f$$

such that

$$\|\theta(\cdot,t)\|_{L^2} \le C(1+t)^{-\frac{1}{2\alpha}}. \tag{3.8}$$

Proof. The arguments of Theorem 3.1 work here and we will point out only the difference. It is easy to see that

$$\|\theta(t)\|_{L^2} \le \|\theta_0\|_{L^2} + \int_0^t \|f(\tau)\|_{L^2} d\tau \le C$$

by energy estimates. When the force f is present, the estimate for $\widehat{\theta}$ is given by

$$|\widehat{\theta}(\xi,t)| \le e^{-|\xi|^{2\alpha}t}|\widehat{\theta_0}| + \int_0^t e^{-|\xi|^{2\alpha}(t-\tau)}\left[\widehat{f} + |\xi|\|\theta\|_{L^2}^2\right] d\tau.$$

Then the procedures of the proof of Theorem 3.1 can be repeated and the assumptions (3.7) are sufficient in establishing (3.8). □

4. Large time approximation. In this section we intend to understand the higher-order correction in the large time approximation of the solution θ to the nonlinear equation by the solution Θ to the linear equation. The approach is to study the difference and the ratio

$$\|\theta(\cdot,t) - \Theta(\cdot,t)\|_{L^2}, \qquad \frac{\|\theta(\cdot,t)\|_{L^2}}{\|\Theta(\cdot,t)\|_{L^2}}.$$

We start with some estimates for the linear equation. The solution of the linear equation on $\mathbb{R}^n$

$$\partial_t\theta + \Lambda^{2\alpha}\theta = 0, \qquad \theta\,|_{t=0} = \theta_0 \tag{4.1}$$

is given by

$$\Theta(t) = k_t^\alpha * \theta_0, \tag{4.2}$$

where the kernel k_t^α is defined by its Fourier transform

$$\widehat{k}_t^\alpha(\xi) = e^{-|\xi|^{2\alpha}t}. \tag{4.3}$$

PROPOSITION 4.1. *Assume that $\alpha > 0$ and the initial data $\theta_0 \in L^1(\mathbb{R}^n)$. Then we have*

$$\lim_{t\to\infty} t^{\frac{n}{2\alpha}}\|\Theta(\cdot,t)\|_{L^2}^2 = A(n,\alpha)\left[\int_{\mathbb{R}^n}\theta_0(x)dx\right]^2, \tag{4.4}$$

$$\lim_{t\to\infty} t^{\frac{n+2}{2\alpha}}\|\nabla\Theta(\cdot,t)\|_{L^2}^2 = B(n,\alpha)\left[\int_{\mathbb{R}^n}\theta_0(x)dx\right]^2, \tag{4.5}$$

where the constants $A(n,\alpha)=\int_{\mathbb{R}^n} e^{-2|\eta|^2}d\eta$ *and* $B(n,\alpha)=\int_{\mathbb{R}^n}|\eta|^2e^{-2|\eta|^2}d\eta$.

Especially for $n=2$, *the* L^2 *decay rates of* Θ *and* $\nabla\Theta$ *are* $t^{-\frac{1}{2\alpha}}$ *and* $t^{-\frac{1}{\alpha}}$, *respectively.*

Proof. We first prove (4.4). By Plancherel's theorem,

$$\lim_{t\to\infty} t^{\frac{n}{2\alpha}}\|\Theta(\cdot,t)\|_{L^2}^2=\lim_{t\to\infty} t^{\frac{n}{2\alpha}}\|\widehat{\Theta}(\cdot,t)\|_{L^2}^2$$

$$=\lim_{t\to\infty} t^{\frac{n}{2\alpha}}\int_{\mathbb{R}^n} e^{-2|\xi|^{2\alpha}t}|\widehat{\theta}_0|^2(\xi)d\xi=\lim_{t\to\infty}\int_{\mathbb{R}^n} e^{-2|\eta|^2}|\widehat{\theta}_0|^2(\eta t^{-\frac{1}{2\alpha}})d\eta.$$

Since for any $t\in[0,\infty)$

$$\int_{\mathbb{R}^n} e^{-2|\eta|^2}|\widehat{\theta}_0|^2(\eta t^{-\frac{1}{2\alpha}})d\eta$$

$$\le\|\widehat{\theta}_0\|_{L^\infty}^2\int_{\mathbb{R}^n} e^{-2|\eta|^2}d\eta\le A(n,\alpha)\|\theta_0\|_{L^1}^2,$$

we can apply the dominated convergence theorem, which leads to (4.4).

The proof of (4.5) is similar to that of (4.4). We have

$$\lim_{t\to\infty} t^{\frac{n+2}{\alpha}}\|\nabla\Theta(\cdot,t)\|_{L^2}^2=\lim_{t\to\infty} t^{\frac{n+2}{\alpha}}\int_{\mathbb{R}^n}|\xi|^2e^{-2|\xi|^{2\alpha}t}|\widehat{\theta}_0|^2(\xi)d\xi$$

$$=\lim_{t\to\infty}\int_{\mathbb{R}^n}|\eta|^2e^{-2|\eta|^2}|\widehat{\theta}_0|^2(\eta t^{-\frac{1}{2\alpha}})d\eta=B(n,\alpha)\left[\int_{\mathbb{R}^n}\theta_0(x)dx\right]^2.\qquad\square$$

PROPOSITION 4.2. *Let* $\alpha\in(0,1]$ *and* $\theta_0\in L^2(\mathbb{R}^2)$. *Then the solution* Θ *of* (4.1) *satisfies*

$$\|\nabla\Theta(t)\|_{L^\infty(\mathbb{R}^2)}\le Ct^{-\frac{1}{\alpha}},$$

where the constant C *depends only on the* L^2 *norm of* θ_0.

Proof. We have by (4.2) and (4.3)

$$\|\nabla\Theta\|_{L^\infty}\le\int_{\mathbb{R}^2}|\xi||\widehat{\Theta}(\xi)|d\xi=\int_{\mathbb{R}^2}|\xi|e^{-|\xi|^{2\alpha}t}|\widehat{\theta}_0(\xi)|d\xi$$

$$\le\|\theta_0\|_{L^2}\left(\int_{\mathbb{R}^2}|\xi|^2e^{-2|\xi|^{2\alpha}t}d\xi\right)^{\frac{1}{2}}\le C\left(\int_0^\infty r^3e^{-2r^{2\alpha}t}dr\right)^{\frac{1}{2}}\le Ct^{-\frac{1}{\alpha}},$$

where the constant C depends only on the L^2 norm of θ_0. $\square$

THEOREM 4.3. *Let* $\alpha\in(0,1]$ *and* $\theta_0\in L^1(\mathbb{R}^2)\cap L^2(\mathbb{R}^2)$. *Then the difference* $\theta-\Theta$ *between a weak solution* θ *of the QGS equation and the solution* Θ *of the linear QGS equation with the data* θ_0 *satisfies*

$$\|\theta(t)-\Theta(t)\|_{L^2(\mathbb{R}^2)}\le C(1+t)^{\frac{1}{2}-\frac{1}{\alpha}},$$

where the constant C *depends only on the* L^1 *and* L^2 *norms of* θ_0.

Remark 4.4. By comparing the rates in Theorems 3.1 and 4.3, we see that $\theta - \Theta$ decays faster than θ does for large time for $\alpha < 1$.

Proof. We will present only a formal argument. The justification process can be done similarly as in the proof of Theorem 3.1. The difference $w = \theta - \Theta$ satisfies

$$\partial_t w + \Lambda^{2\alpha} w = -u \cdot \nabla \theta. \tag{4.6}$$

Taking the scalar product of (4.6) with w and using the fact that

$$\int_{\mathbb{R}^2} (u \cdot \nabla \theta) \theta dx = 0,$$

we obtain

$$\frac{d}{dt} \int_{\mathbb{R}^2} |w|^2 + 2 \int_{\mathbb{R}^2} |\Lambda^\alpha w|^2 = \int_{\mathbb{R}^2} \Theta (u \cdot \nabla \theta) dx.$$

Using the results of Proposition 4.2 and Theorem 3.1, we bound the right-hand term by

$$\left| \int_{\mathbb{R}^2} \Theta (u \cdot \nabla \theta) dx \right| \le \|\nabla \Theta\|_{L^\infty} \|\theta\|_{L^2}^2 \le C(1+t)^{-\frac{2}{\alpha}}.$$

As in the proof of Theorem 3.1,

$$\frac{d}{dt} \int_{\mathbb{R}^2} |\widehat{w}|^2 + 2g^{2\alpha}(t) \int_{\mathbb{R}^2} |\widehat{w}|^2 \le 2g^{2\alpha}(t) \int_{|\xi| \le g(t)} |\widehat{w}|^2 + C(1+t)^{-\frac{2}{\alpha}}, \tag{4.7}$$

where $g(t)$ remains to be decided.

We need an estimate of $\widehat{w}$, which can be obtained by taking the Fourier transform of (4.6) and proceeding as in Lemma 3.3. By Theorem 3.1 and noticing $\alpha \le 1$,

$$|w(\xi, t)| \le |\xi| \int_0^t \|\theta(\tau)\|_{L^2}^2 d\tau \le |\xi| \int_0^t (1+\tau)^{-\frac{1}{\alpha}} d\tau \le C|\xi|.$$

Taking $g^{2\alpha} = \frac{\beta}{2(1+t)}$, we obtain, by integrating (4.7),

$$(1+t)^\beta \int_{\mathbb{R}^2} |\widehat{w}|^2 \le C \left[\int_0^t (1+\tau)^{\beta - \frac{2}{\alpha}} d\tau + \int_0^t (1+\tau)^\beta g^4(\tau) d\tau \right].$$

Therefore,

$$\|w\|_{L^2}^2 \le C(1+t)^{1-\frac{2}{\alpha}}.$$

This completes the proof of Theorem 4.3. □

We can consider lower bounds for the decay of θ with the aid of Theorem 4.3. It is easy to see that Θ can decay exponentially fast. For example, if $\widehat{\theta_0} = 0$ for $|\xi| \le \gamma$, then

$$\|\Theta(t)\|_{L^2}^2 = \int e^{-2|\xi|^{2\alpha} t} |\widehat{\theta_0}(\xi)|^2 \le \|\theta_0\|_{L^2}^2 e^{-2\gamma^2 t}.$$

However, for those θ_0 satisfying

$$|\widehat{\theta_0}(\xi)| \ge \lambda \qquad \text{for} \qquad |\xi| \le \gamma, \tag{4.8}$$

we have the following.

PROPOSITION 4.5. *Let $\alpha \in (0,1]$ and $\theta_0 \in L^2(\mathbb{R}^2)$ satisfy (4.8). Then if Θ is a solution of the linear QGS equation,*

$$\|\Theta(t)\|_{L^2(\mathbb{R}^2)} \geq C(1+t)^{-\frac{1}{2\alpha}},$$

where C is a constant depending only on λ, γ, and the L^2 norm of θ_0.

As a corollary of Theorem 4.3 and Proposition 4.5, we have the following.

THEOREM 4.6. *Let $\alpha \in (0,1]$ and $\theta_0 \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ satisfy (4.8). Then for a weak solution θ of the QGS equation with data θ_0,*

$$\|\theta(\cdot,t)\|_{L^2(\mathbb{R}^2)} \geq C(1+t)^{-\frac{1}{2\alpha}},$$

where C depends on λ, γ, and the L^1 and L^2 norms of θ_0.

The following theorem reveals more detailed aspects of the higher-order correction.

THEOREM 4.7. *Let $\alpha \in (\frac{1}{2}, 1]$ and $\delta > 0$ such that $2\alpha - 1 - \delta \geq 0$. Assume that θ is a weak solution of the 2D QGS equation*

$$\partial_t \theta + u \cdot \nabla \theta + \Lambda^{2\alpha}\theta = 0$$

with initial data $\theta_0 \in L^1(\mathbb{R}^2) \cap L^2(\mathbb{R}^2)$ and that satisfies

$$\|\Lambda^{2-2\alpha+\delta}\theta(\cdot,t)\|_{L^2} \leq Ct^{-\epsilon} \tag{4.9}$$

for some constant C and $\epsilon > 0$. Let Θ be the solution of the linear equation with the same initial data θ_0. Then

$$\frac{\|\theta(\cdot,t)\|_{L^2}}{\|\Theta(\cdot,t)\|_{L^2}} = 1 + O(t^{-\min\{\frac{1}{2\alpha},\epsilon\}}), \tag{4.10}$$

$$t^{\frac{1}{2\alpha}+\min\{\frac{1}{2\alpha},\epsilon\}}\|\theta(\cdot,t) - \Theta(\cdot,t)\|_{L^2} = O(1). \tag{4.11}$$

Proof. By taking the Fourier transform of the equation for θ

$$\partial_t \theta + \Lambda^{2\alpha}\theta = -u \cdot \nabla \theta,$$

we obtain

$$\widehat{\theta}(\xi,t) = e^{-|\xi|^{2\alpha}t}\widehat{\theta_0}(\xi) - \int_0^t e^{-|\xi|^{2\alpha}(t-s)}\widehat{u \cdot \nabla \theta}(\xi,s)ds.$$

Then the ratio

$$\frac{\|\theta(\cdot,t)\|_{L^2}^2}{\|\Theta(\cdot,t)\|_{L^2}^2} = \frac{\|\widehat{\theta}(\cdot,t)\|_{L^2}^2}{\|\widehat{\Theta}(\cdot,t)\|_{L^2}^2} = 1 + 2\mathcal{J}(t) + \mathcal{J}^2(t),$$

where $\mathcal{J}$ is given by

$$\mathcal{J}(t) = \frac{\int_{\mathbb{R}^2}\left|\int_0^t e^{-|\xi|^{2\alpha}(t-s)}\widehat{u \cdot \nabla \theta}(\xi,s)ds\right|^2 d\xi}{\|\Theta\|_{L^2}^2}.$$

To prove (4.10), it suffices to show that

$$\mathcal{J}(t) = O(t^{-\min\{\frac{1}{2\alpha},\epsilon\}}).$$

The difference $w = \theta - \Theta$ satisfies

$$\partial_t w + \Lambda^{2\alpha} w = -u \cdot \nabla \theta.$$

Since $w(x,0) = \theta(x,0) - \Theta(x,0) = 0$,

$$\widehat{w}(\xi,t) = -\int_0^t e^{-|\xi|^{2\alpha}(t-s)} \widehat{u \cdot \nabla \theta}(\xi,s) ds,$$

$$\|w(\cdot,t)\|_{L^2}^2 = \|\widehat{w}(\cdot,t)\|_{L^2}^2 = \int_{\mathbb{R}^2} \left| \int_0^t e^{-|\xi|^{2\alpha}(t-s)} \widehat{u \cdot \nabla \theta}(\xi,s) ds \right|^2 d\xi.$$

Thus, to prove (4.10) and (4.11), we need only to estimate the integral

$$I \equiv \int_{\mathbb{R}^2} \left| \int_0^t e^{-|\xi|^{2\alpha}(t-s)} \widehat{u \cdot \nabla \theta}(\xi,s) ds \right|^2 d\xi.$$

To this end, we divide the integral I into the following two parts:

$$II = \int_{\mathbb{R}^2} \left| \int_0^{t/2} \cdots \right|^2 d\xi \quad \text{and} \quad III = \int_{\mathbb{R}^2} \left| \int_{t/2}^t \cdots \right|^2 d\xi.$$

Since $\nabla \cdot u = 0$, $\widehat{u \cdot \nabla \theta} = i\xi \cdot \widehat{u\theta}$ and we obtain, by setting $\eta = t^{\frac{1}{2\alpha}} \xi$,

$$II = t^{-\frac{2}{\alpha}} \int_{\mathbb{R}^2} e^{-2|\eta|^{2\alpha}} \left| \int_0^{t/2} e^{\frac{s}{t}|\eta|^2} \eta \cdot \widehat{u\theta}(\eta t^{-\frac{1}{2\alpha}}, s) ds \right|^2 d\eta.$$

Since

$$\|\widehat{u\theta}(\cdot,s)\|_{L^\infty} \le \|u\theta(\cdot,s)\|_{L^1} \le \|u(\cdot,s)\|_{L^2} \|\theta(\cdot,s)\|_{L^2} \le C \|\theta(\cdot,s)\|_{L^2}^2,$$

we have the bound

$$II \le C t^{-\frac{2}{\alpha}} \int_{\mathbb{R}^2} |\eta|^2 e^{-\frac{7}{4}|\eta|^2} \left[\int_0^\infty \|\theta(\cdot,s)\|_{L^2}^2 ds \right]^2 d\eta$$

$$\le C t^{-\frac{2}{\alpha}} \left(\int_{\mathbb{R}^2} |\eta|^2 e^{-\frac{7}{4}|\eta|^2} d\eta \right) \|\theta\|_{L^2([0,\infty);L^2)}^4.$$

The estimate of III seems tricky. Intuitively, the idea is to split the whole derivative ∇ into two fractional parts $\Lambda^{2\alpha-1-\delta}$ and $\Lambda^{2-2\alpha+\delta}$:

$$III = \int_{\mathbb{R}^2} \left| \int_{t/2}^t e^{-|\xi|^{2\alpha}(t-s)} \widehat{u \cdot \nabla \theta}(\xi,s) ds \right|^2 d\xi$$

$$\le \int_{\mathbb{R}^2} e^{-2|\xi|^{2\alpha} t} |\xi|^{2(2\alpha-1-\delta)} \sup_{t/2 \le s \le t} |u \cdot \widehat{\Lambda^{2-2\alpha+\delta}\theta}|^2(\xi,s) \left[\int_{t/2}^t e^{s|\xi|^{2\alpha}} ds \right]^2 d\xi.$$

Using the assumption (4.9), we obtain

$$|u \cdot \widehat{\Lambda^{2-2\alpha+\delta}\theta}|(\xi,s) \le \|u \cdot \widehat{\Lambda^{2-2\alpha+\delta}\theta}(\cdot,s)\|_{L^\infty}$$

$$\le \|(u \cdot \Lambda^{2-2\alpha+\delta}\theta)(\cdot,s)\|_{L^1} \le C\|u(\cdot,s)\|_{L^2}\|\Lambda^{2-2\alpha+\delta}\theta(\cdot,s)\|_{L^2} \le Cs^{-\frac{1}{2\alpha}-\epsilon},$$

where C is a constant. Therefore

$$III \le Ct^{-\frac{1}{\alpha}-2\epsilon}\int_{\mathbb{R}^2} |\xi|^{-2-\delta}(1-e^{-\frac{1}{2}|\xi|^{2\alpha}t})^2 d\xi \le Ct^{-\frac{1}{\alpha}-2\epsilon}.$$

Combining the estimates for II and III, we conclude that

$$I \equiv \int_{\mathbb{R}^2}\left|\int_0^t e^{-|\xi|^{2\alpha}(t-s)}\widehat{u\cdot\nabla\theta}(\xi,s)ds\right|^2 d\xi \le Ct^{-\frac{1}{\alpha}-\min\{\frac{1}{\alpha},2\epsilon\}},$$

and (4.10), (4.11) are therefore established. □

Acknowledgment. We thank Professor Jerry Bona for discussions.

REFERENCES

[1] C. Amick, J. Bona, and M. Schonbek, *Decay of solutions of some nonlinear wave equations*, J. Differential Equations, 81 (1989), pp. 1–49.
[2] L. Caffarelli, R. Kohn, and L. Nirenberg, *Partial regularity of suitable weak solutions of the Navier-Stokes equations*, Comm. Pure Appl. Math., 35 (1982), pp. 771–831.
[3] P. Constantin and C. Foias, *Navier-Stokes Equations*, University of Chicago Press, Chicago, IL, 1988.
[4] P. Constantin, A. Majda, and E. Tabak, *Formation of strong fronts in the 2-D quasi-geostrophic thermal active scalar*, Nonlinearity, 7 (1994), pp. 1495–1533.
[5] I. Held, R. Pierrehumbert, S. Garner, and K. Swanson, *Surface quasi-geostrophic dynamics*, J. Fluid Mech., 282 (1995), pp. 1–20.
[6] P. L. Lions, *Mathematical Topics in Fluid Mechanics*, Vol. 1, The Clarendon Press, Oxford, 1996.
[7] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer-Verlag, New York, 1987.
[8] S. Resnick, *Dynamical Problems in Non-linear Advective Partial Differential Equations*, Ph.D. thesis, University of Chicago, Chicago, IL, 1995.
[9] M. Schonbek, *L^2 decay for weak solutions of the Navier-Stokes equations*, Arch. Rational Mech. Anal., 88 (1985), pp. 209–222.
[10] M. Schonbek, *Large time behavior of solutions to the Navier-Stokes equations*, Comm. Partial Differential Equations, 11 (1986), pp. 733–763.
[11] E. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[12] M. Taylor, *Pseudodifferential Operators and Nonlinear PDE*, Birkhäuser, Boston, 1991.
[13] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1984.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 949–962

LARGE EIGENVALUES AND TRACE FORMULAS FOR MATRIX STURM–LIOUVILLE PROBLEMS*

ROBERT CARLSON†

Abstract. This work describes the location of large eigenvalues for Sturm–Liouville operators with matrix coefficients. Eigenvalue asymptotics of arbitrary order are characterized for operators with smooth coefficients and "elementary" boundary conditions. In addition, an effective method is given for computing traces of these operators. These techniques are employed to verify a previously conjectured trace formula.

Key words. trace formulas, eigenvalue asymptotics, Sturm–Liouville problems

AMS subject classifications. 34B24, 34L20

PII. S0036141098340417

1. Introduction. Although ordinary differential operators do not have a trace in the usual sense, it was observed some time ago [2] that one could often make sense of $\sum_n(\mu_n-\nu_n)$, where $\{\mu_n\}$ and $\{\nu_n\}$ are the eigenvalues of two differential operators. Moreover, these traces are often given by explicit expressions in the coefficients of the operators. Beyond their aesthetic appeal, trace formulas play an important role in inverse spectral theory [4, 6, 10, 12].

Several authors [5, 11] have raised the problem of establishing and analyzing trace formulas for the one-dimensional matrix Schrödinger equation

$$-Y'' + Q(x)Y = \lambda Y, \quad Y \in \mathbf{C}^K, \quad \lambda \in \mathbf{C}, \tag{1a}$$

subject to certain boundary conditions. In particular, Papanicolaou [11] established a trace formula for the Dirichlet boundary conditions $Y(0) = 0 = Y(1)$ when the $K \times K$ matrix $Q(x)$ is real symmetric, subject to a conjectured asymptotic behavior of the eigenvalue sequence.

In this work we will show how to establish eigenvalue estimates of any order using classical asymptotic expansions for solutions of (1a). Eigenvalues for (1a) subject to certain separated boundary conditions are identified with the vanishing of $\det \mathcal{T}(\lambda)$, where $\mathcal{T}(\lambda)$ is an entire $K \times K$ matrix-valued function. A novel feature of our approach is the use of perturbation theory to analyze the location of the eigenvalues of $\mathcal{T}(\lambda)$ or a closely related matrix. This approach appears to offer advantages over the direct consideration of the equation $\det \mathcal{T}(\lambda) = 0$. We also extend some old ideas for computing (higher order) traces. These techniques downplay the role of self-adjoint operators.

2. Regularized traces. It is common to use infinite product representations [1] to relate the eigenvalue sequence for an ordinary differential operator to an entire function having the eigenvalues as zeroes. The first lemma considers the computation of power sums

$$\sum_{n=1}^{\infty}(\nu_n^k - \mu_n^k)$$

*Received by the editors June 1, 1998; accepted for publication October 26, 1998; published electronically June 29, 1999.
http://www.siam.org/journals/sima/30-5/34041.html
†Department of Mathematics, University of Colorado at Colorado Springs, Colorado Springs, CO, 80933 (carlson@math.uccs.edu).

based on the asymptotic behavior of such entire functions. A similar treatment is given in [8, pp. 84–88], where the ideas are attributed to F. Schäfke.

LEMMA 2.1. *Suppose that $f(\lambda)$ and $g(\lambda)$ are entire functions not vanishing on a real half line $(-\infty, b)$, which are given by absolutely convergent products*

$$f(\lambda) = \prod_{n=1}^{\infty} \left(1 - \frac{\lambda}{\mu_n}\right), \quad g(\lambda) = \prod_{n=1}^{\infty} \left(1 - \frac{\lambda}{\nu_n}\right).$$

(*For notational convenience we assume that $f(0)g(0) \neq 0$.*)

If all but finitely many μ_n, ν_n lie in the right half plane, and

$$\sum |\mu_n - \nu_n|[|\mu_n|^{K-1-\epsilon} + |\nu_n|^{K-1-\epsilon}] < \infty, \quad 0 \leq \epsilon < 1,$$

for an integer $K \geq 2$, then as $\lambda \to -\infty$ along the real axis,

$$\log\left(\frac{f(\lambda)}{g(\lambda)}\right) = C + \sum_{k=1}^{K-1} \Big[\frac{1}{k\lambda^k} \sum_{n=1}^{\infty} (\nu_n^k - \mu_n^k)\Big] + O(\lambda^{-K+\epsilon}).$$

Proof. For $\lambda \in (-\infty, b)$ define

$$\begin{aligned} h(\lambda) &= \partial_\lambda \log\left(\frac{f}{g}\right) = \partial_\lambda \log\left(\prod \frac{\lambda - \mu_n}{\lambda - \nu_n}\right) \\ &= \sum_n \left(\frac{1}{\lambda - \mu_n} - \frac{1}{\lambda - \nu_n}\right) = \sum_n \frac{1}{\lambda}\left(\frac{1}{1 - \mu_n/\lambda} - \frac{1}{1 - \nu_n/\lambda}\right). \end{aligned}$$

Use of the identity $1/(1-x) = 1 + x + \cdots + x^{K-1} + x^K/(1-x)$ gives

$$h(\lambda) = \sum_n \left[\frac{1}{\lambda} \sum_{k=1}^{K-1} \frac{\mu_n^k - \nu_n^k}{\lambda^k} + \frac{1}{\lambda}\left[\frac{(\mu_n/\lambda)^K}{1 - \mu_n/\lambda} - \frac{(\nu_n/\lambda)^K}{1 - \nu_n/\lambda}\right]\right].$$

A rearrangement of these sums is justified if the sums

$$\sum_n |\mu_n^k - \nu_n^k|, \quad k = 1, \ldots, K-1, \tag{2a}$$

and

$$\sum_n \left|\frac{\mu_n^K}{1 - \mu_n/\lambda} - \frac{\nu_n^K}{1 - \nu_n/\lambda}\right| \tag{2b}$$

are convergent. Using the fundamental theorem of calculus to write the difference $\mu_n^k - \nu_n^k$ as a contour integral over the line segment joining the endpoints, we find

$$\sum_n |\mu_n^k - \nu_n^k| = \sum_n \left|\int_{\nu_n}^{\mu_n} kz^{k-1}\, dz\right| \leq k \sum_n |\mu_n - \nu_n|[|\mu_n|^{k-1} + |\nu_n|^{k-1}].$$

There are only finitely many μ_n, ν_n inside the unit disk. Except for this finite collection we have

$$|\mu_n - \nu_n|[|\mu_n|^{k-1} + |\nu_n|^{k-1}] \leq |\mu_n - \nu_n|[|\mu_n|^{K-2} + |\nu_n|^{K-2}], \quad 1 \leq k \leq K-1.$$

Thus the sums in (2a) are absolutely convergent.

As for (2b) the same technique gives

$$\sum_n \left| \frac{\mu_n^K}{1-\mu_n/\lambda} - \frac{\nu_n^K}{1-\nu_n/\lambda} \right| = \sum_n \left| \lambda \int_{\nu_n}^{\mu_n} \frac{Kz^{K-1}}{\lambda - z} + \frac{z^K}{(\lambda - z)^2}\, dz \right|. \tag{2c}$$

If μ_n, ν_n have positive real parts, $0 \le \epsilon < 1$, and $\lambda \in (-\infty, 0)$, then

$$|\lambda - z|^{-1} \le |\lambda|^{-1+\epsilon} |z|^{-\epsilon}, \quad |\lambda - z|^{-2} \le |\lambda|^{-1+\epsilon} |z|^{-1-\epsilon}.$$

Thus after dropping finitely many terms, the right-hand side of (2c) is bounded by

$$[K+1]|\lambda|^{\epsilon} \sum_n |\mu_n - \nu_n| [|\mu_n|^{K-1-\epsilon} + |\nu_n|^{K-1-\epsilon}].$$

Rearrangement now gives

$$h(\lambda) = \partial_\lambda \log\left(\frac{f}{g}\right) = \sum_{k=1}^{K-1} \Big[\lambda^{-k-1} \sum_n (\mu_n^k - \nu_n^k)\Big] + O(\lambda^{-K-1+\epsilon}),$$

and integration gives the desired result. □

3. Asymptotic expansions for solutions of (1a). Estimates for solutions to the initial value problem for (1a) may be established using familiar techniques from the scalar case. With the exception of some minor notational changes, and the rearrangement of some terms which do not commute in the vector case, the form of our estimates is very close to that of [3]. Since the arguments from the scalar case are also applicable, the reader may consult this reference for the proof of Lemmas 3.1 and 3.2. Related problems are treated with a somewhat different approach in [9, pp. 50–100].

It will be helpful to establish some notational conventions. For $\lambda \in \mathbf{C}$ let $\omega = \sqrt{\lambda}$, the root chosen continuously for $-\pi \le \arg(\lambda) < \pi$ and positive for $\lambda > 0$. The imaginary part of ω is denoted by $\Im\omega$. An element $Y \in \mathbf{C}^K$ will have the usual norm

$$|Y| = \left[\sum_{k=1}^K |y_k|^2\right]^{1/2}, \quad Y = \begin{pmatrix} y_1 \\ \vdots \\ y_K \end{pmatrix},$$

and a $K \times K$ matrix Q will have the operator norm

$$\|Q\| = \sup_{|Y|=1} |QY|.$$

The $K \times K$ identity matrix is I_K, and the zero matrix is 0_K.

A brief discussion will introduce the ideas. The model equation $-Y'' = \lambda Y$ has a basis of $2K$ solutions which are the columns of the $K \times K$ diagonal matrix-valued functions $\cos(\omega x) I_K$, $\omega^{-1} \sin(\omega x) I_K$. Adapting the variation of parameters formula to this setting, a solution of (1a) satisfying $Y(0,\lambda) = \alpha$, $Y'(0,\lambda) = \beta$, with $\alpha, \beta \in \mathbf{C}^K$, may be written as a solution of the integral equation

$$Y(x,\lambda) = \cos(\omega x)\alpha + \omega^{-1} \sin(\omega x)\beta + \omega^{-1} \int_0^x \sin(\omega[x-t]) Q(t) Y(t,\lambda)\, dt. \tag{3a}$$

Differentiation with respect to x gives

$$Y'(x,\lambda) = -\omega \sin(\omega x)\alpha + \cos(\omega x)\beta + \int_0^x \cos(\omega[x-t])Q(t)Y(t,\lambda)\ dt.$$

When $Q(x)$ is sufficiently differentiable an expansion of $Y(x,\lambda)$, together with error estimates, may be obtained by an iteration scheme based on (3a) and integration by parts. For $J \geq 1$ let C^J denote the Banach space of $K \times K$ matrix-valued functions $Q(x)$ whose components have J continuous derivatives on $[0,1]$. We can equip C^J with the norm

$$\|Q\|_J = \max \|Q^{(j)}(x)\|, \quad 0 \leq j \leq J, \quad 0 \leq x \leq 1.$$

For notational convenience define $\mathcal{A}_j(x) = Q(x)A_j(x)$ and $\mathcal{B}_j(x) = Q(x)B_j(x)$.

LEMMA 3.1. *Suppose that $Q \in C^J$, and $Y(x,\lambda)$ is the solution of* (1a) *satisfying $Y(0,\lambda) = \alpha$ and $Y'(0,\lambda) = \beta$, with $|\alpha| + |\beta| \leq 1$, $0 \leq x \leq 1$, and $|\omega| \geq 1$. Then there are $\mathbf{C}^K$ valued functions $A_j(x)$ and $B_j(x)$ such that*

$$\left| Y(x,\lambda) - \sum_{j=0}^{J} \omega^{-j}[\cos(\omega x)A_j(x) + \sin(\omega x)B_j(x)] \right| = O(\omega^{-J-1}e^{|\Im\omega|x}).$$

The coefficients $A_j(x), B_j(x)$ satisfy

$$A_0(x) = \alpha, \quad A_1(x) = 0,$$
$$A_2(x) = 2^{-2}[Q(x) - Q(0)]\alpha - 2^{-1}\int_0^x Q(t)\left[\beta + 2^{-1}\int_0^t Q(s)\alpha\ ds\right] dt,$$
$$B_0(x) = 0, \quad B_1(x) = \beta + 2^{-1}\int_0^x Q(t)\alpha\ dt, \quad B_2(x) = 0,$$

and for $j \geq 3$, the coefficients satisfy the recursion relations

$$A_j(x) = \sum_{k=0}^{\lfloor (j-2)/2 \rfloor} (-1)^k 2^{-2k-2}[\mathcal{A}_{j-2k-2}^{(2k)}(x) - \mathcal{A}_{j-2k-2}^{(2k)}(0)]$$
$$- 2^{-1}\int_0^x \mathcal{B}_{j-1}(t)\ dt + \sum_{k=0}^{\lfloor (j-3)/2 \rfloor} (-1)^k 2^{-2k-3}[\mathcal{B}_{j-2k-3}^{(2k+1)}(x) - \mathcal{B}_{j-2k-3}^{(2k+1)}(0)],$$
$$B_j(x) = -\sum_{k=0}^{\lfloor (j-3)/2 \rfloor} (-1)^k 2^{-2k-3}[\mathcal{A}_{j-2k-3}^{(2k+1)}(x) + \mathcal{A}_{j-2k-3}^{(2k+1)}(0)]$$
$$+ 2^{-1}\int_0^x \mathcal{A}_{j-1}(t)\ dt + \sum_{k=0}^{\lfloor (j-2)/2 \rfloor} (-1)^k 2^{-2k-2}[\mathcal{B}_{j-2k-2}^{(2k)}(x) + \mathcal{B}_{j-2k-2}^{(2k)}(0)].$$

The function $Y'(x,\lambda)$ has a similar expansion obtained from that of $Y(x,\lambda)$ by termwise differention.

LEMMA 3.2. *Under the hypotheses of Lemma* 3.1 *there are $\mathbf{C}^K$ valued functions $C_j(x)$ and $D_j(x)$ such that*

$$\left| Y'(x,\lambda) - \sum_{j=-1}^{J-1} \omega^{-j}[\cos(\omega x)C_j(x) + \sin(\omega x)D_j(x)] \right| = O(\omega^{-J}e^{|\Im\omega|x}).$$

The coefficients $C_j(x), D_j(x)$ *satisfy* $C_{-1}(x) = 0$, $D_{-1}(x) = -\alpha$, *and*

$$C_j(x) = A_j'(x) + B_{j+1}(x), \quad D_j(x) = B_j'(x) - A_{j+1}(x), \quad j = 0, \dots, J-1.$$

In particular, one has the following expression:

$$\begin{aligned}
Y'(x,\lambda) &= -\omega \sin(\omega x)\alpha + \cos(\omega x)\beta + 2^{-1}\cos(\omega x)\int_0^x Q(t)\ dt\ \alpha \\
&\quad + 2^{-2}\omega^{-1}\sin(\omega x)[Q(x)+Q(0)]\alpha + 2^{-1}\omega^{-1}\sin(\omega x)\int_0^x Q(t)\ dt\ \beta \\
&\quad + 2^{-2}\omega^{-1}\sin(\omega x)\int_0^x Q(t)\int_0^t Q(s)\ ds\ dt\ \alpha \\
&\quad + 2^{-3}\omega^{-2}\cos(\omega x)[Q'(x)-Q'(0)]\alpha - 2^{-2}\omega^{-2}\cos(\omega x)[Q(x)-Q(0)]\beta \\
&\quad - 2^{-3}\omega^{-2}\cos(\omega x)Q(x)\int_0^x Q(t)\ dt\ \alpha \\
&\quad + 2^{-3}\omega^{-2}\cos(\omega x)\int_0^x Q(t)[Q(t)-Q(0)]\ dt\ \alpha \\
&\quad - 2^{-2}\omega^{-2}\cos(\omega x)\int_0^x Q(t)\int_0^t Q(s)\ ds\ dt\ \beta \\
&\quad - 2^{-3}\omega^{-2}\cos(\omega x)\int_0^x Q(t)\int_0^t Q(s)\int_0^s Q(u)\ du\ ds\ dt\ \alpha \\
&\quad + O(\omega^{-3}\exp(|\Im(\omega)|)).
\end{aligned}$$

It will be convenient to introduce the $K \times K$ matrix solutions $C(x,\lambda), S(x,\lambda)$ of (1a) which satisfy

$$\begin{aligned}
C(0,\lambda) &= I_K, \quad C'(0,\lambda) = 0_K, \\
S(0,\lambda) &= 0_K, \quad S'(0,\lambda) = I_K.
\end{aligned}$$

The columns of these matrices are a basis of solutions to (1a). The asymptotic expansions for $Y(x,\lambda)$ and $Y'(x,\lambda)$ specialize to give expansions for the matrix functions $C(1,\lambda)$, $C'(1,\lambda)$, $S(1,\lambda)$, and $S'(1,\lambda)$.

Defining $Q_0 = \int_0^1 Q(t)\ dt$, the following explicit formulas will be needed:

(3b)

$$\begin{aligned}
C(1,\lambda) &= \cos(\omega)I_K + 2^{-1}\omega^{-1}\sin(\omega)Q_0 \\
&\quad + 2^{-2}\omega^{-2}\cos(\omega)\Big[Q(1)-Q(0)-\int_0^1 Q(t)\int_0^t Q(s)\ ds\ dt\Big] + O(\omega^{-3}e^{|\Im\omega|}). \\
S(1,\lambda) &= \omega^{-1}\sin(\omega)I_K - 2^{-1}\omega^{-2}\cos(\omega)Q_0 \\
&\quad + 2^{-2}\omega^{-3}\sin(\omega)\Big[Q(1)+Q(0)-\int_0^1 Q(t)\int_0^t Q(s)\ ds\ dt\Big] + O(\omega^{-4}e^{|\Im\omega|}), \\
C'(1,\lambda) &= -\omega\sin(\omega)I_K + 2^{-1}\cos(\omega)Q_0 \\
&\quad + 2^{-2}\omega^{-1}\sin(\omega)\Big[Q(1)+Q(0)+\int_0^1 Q(t)\int_0^t Q(s)\ ds\ dt\Big] + O(\omega^{-2}e^{|\Im\omega|}), \\
S'(1,\lambda) &= \cos(\omega)I_K + 2^{-1}\omega^{-1}\sin(\omega)Q_0 \\
&\quad - 2^{-2}\omega^{-2}\cos(\omega)\left[Q(1)-Q(0)+\int_0^1 Q(t)\int_0^t Q(s)\ ds\ dt\right] + O(\omega^{-3}e^{|\Im\omega|}).
\end{aligned}$$

4. Eigenvalues and traces. Eigenvalue problems given by (1a), together with separated boundary conditions, will be considered next. For simplicity the eigenvalue problem is assumed to have the form

$$\text{(4a)} \qquad \begin{aligned} &-Y'' + Q(x)Y = \lambda Y, \quad Y \in \mathbf{C}^K, \quad \lambda \in \mathbf{C}, \\ &Y_i^{(j(i))}(0) = 0 = Y_i^{(k(i))}(1), \quad i = 1, \dots, K, \quad j(i), k(i) \in \{0, 1\}, \end{aligned}$$

so that for each component either the value of the function or its derivative vanishes at each endpoint. To further facilitate the development, assume that the matrix Q_0 is diagonal, $Q_0 = \operatorname{diag}[q_1, \dots, q_K]$. Each of the expansions (3b) now has diagonal matrix coefficients for the first two terms. Under the weaker assumption that Q_0 is similar to a diagonal matrix, a simple linear change of variables reduces (1a) to the case of diagonal Q_0. Some of the results below would require modifications to account for this transformation.

For each λ there is a K-dimensional space of solutions to the eigenvalue equation satisfying the boundary conditions at 0. A basis for these solutions is given by the columns of the $K \times K$ matrix $T(x, \lambda)$, obtained by selecting K columns from the $K \times 2K$ matrix $(C(x, \lambda)\ S(x, \lambda))$, where the ith boundary condition at 0 dictates the selection of column i if the derivative vanishes and the $(K + i)$th column if the function vanishes. The problem (4a) will have an eigenvalue at λ if and only if some nontrivial linear combination of these basis functions satisfies the boundary conditions at 1.

Define the $K \times K$ matrix $\mathcal{T}(\lambda)$ whose entries $T_{ij}^{(k(i))}(1, \lambda)$ are obtained by applying the ith boundary functional at 1 from (4a) to the jth column of T. The ijth entry of $\mathcal{T}(\lambda)$ is, therefore, the ijth entry of one of the matrices $C(1, \lambda)$, $S(1, \lambda)$, $C'(1, \lambda)$, or $S'(1, \lambda)$. The problem (4a) has an eigenvalue at λ if and only if $\det \mathcal{T}(\lambda) = 0$ or, equivalently, if one of the eigenvalues of $\mathcal{T}(\lambda)$ is 0.

The eigenvalues of (4a) will be compared to the eigenvalues of $-D^2 + Q_0$ with the same boundary conditions. Since Q_0 is diagonal and the boundary conditions respect the decoupling of the system of equations, we may partition the eigenvalues into K subsequences $\lambda_{n,k}^0$ which are the eigenvalues of the kth component problem. For each k the subsequence is listed with increasing real parts. The eigenvalues $\lambda_{n,k}^0$ are the roots of one of the functions $\cos(\sqrt{\lambda - q_k})$, $\sin(\sqrt{\lambda - q_k})/\sqrt{\lambda - q_k}$, or $\sqrt{\lambda - q_k}\sin(\sqrt{\lambda - q_k})$, respectively, giving

$$\lambda_{n,k}^0 = \left\{ \begin{matrix} [n - \frac{1}{2}]^2\pi^2 & + & q_k, \\ n^2\pi^2 & + & q_k, \\ [n - 1]^2\pi^2 & + & q_k, \end{matrix} \right\} \quad n = 1, 2, 3, \dots.$$

THEOREM 4.1. *Suppose that Q_0 is a diagonal matrix and $Q(x) \in C^2$. There is a sequence $d_{n,k}$ of disks containing the $\lambda_{n,k}^0$ with radii $\alpha_{n,k} = O(n^{-1})$ such that every disk contains at least one eigenvalue of* (4a) *and every eigenvalue of* (4a) *lies inside a disk.*

If the off-diagonal entries of the matrices $Q(1)$, $Q(0)$, and $\int_0^1 Q(t) \int_0^t Q(s)\, ds\, dt$ vanish, or if the boundary conditions have the form $Y_i^{(j)}(0) = 0 = Y_i^{(k)}(1)$ for $j, k \in \{0, 1\}$ (not depending on i), then this estimate improves to $\alpha_{n,k} = O(n^{-2})$.

Proof. For $\omega \neq 0$ define diagonal matrices Ω_1, Ω_2 as follows. Ω_1 has ith diagonal entry 1 (respectively, ω^{-1}) if the ith boundary condition at 1 requires the function (respectively, derivative) to vanish. Ω_2 has ith diagonal entry ω (respectively, 1) if the

ith boundary condition at 0 requires the function (respectively, derivative) to vanish. Now define $R = \Omega_1 \mathcal{T} \Omega_2$.

The matrix $R = (r_{ij})$ has off-diagonal entries which are $O(\omega^{-2} e^{|\Im\omega|})$. The kth diagonal entry r_{kk} of R has one of the forms

$$r_{kk} = \cos(\omega) + 2^{-1} q_k \omega^{-1} \sin(\omega) + c_k \omega^{-2} \cos(\omega) + O(\omega^{-3} e^{|\Im\omega|})$$

or

$$\pm r_{kk} = \sin(\omega) - 2^{-1} q_k \omega^{-1} \cos(\omega) + c_k \omega^{-2} \cos(\omega) + O(\omega^{-3} e^{|\Im\omega|}).$$

The constant $\pm c_k$, which may be obtained from (3b), is the kth diagonal entry of one of the matrices

$$Q(1) \pm Q(0) \pm \int_0^1 Q(t) \int_0^t Q(s).$$

Since

$$\cos(\sqrt{\lambda - q_k}) = [1 - 2^{-3}\omega^{-2} q_k^2] \cos(\omega) + 2^{-1}\omega^{-1} q_k \sin(\omega) + O(\omega^{-3} \exp(|\Im\omega|)),$$
$$\sin(\sqrt{\lambda - q_k}) = [1 - 2^{-3}\omega^{-2} q_k^2] \sin(\omega) - 2^{-1}\omega^{-1} q_k \cos(\omega) + O(\omega^{-3} \exp(|\Im\omega|)),$$

the diagonal entries of R may be written as $\rho_k + O(\omega^{-3} e^{|\Im\omega|})$, where

$$\rho_k = \left(1 + \omega^{-2}[2^{-3} q_k^2 + c_k]\right) \cos(\sqrt{\lambda - q_k})$$

or

$$\pm\rho_k = \left(1 + \omega^{-2}[2^{-3} q_k^2 + c_k]\right) \sin(\sqrt{\lambda - q_k}).$$

If $\lambda \neq 0$, the matrix $\mathcal{T}(\lambda)$ has an eigenvalue 0 if and only if $R(\lambda)$ does. Write $R = R_0 + R_1$, where R_0 is diagonal with entries ρ_k and $R_1 = O(\omega^{-2} e^{|\Im\omega|})$. If $\Sigma(R)$ denotes the set of eigenvalues of R, then since R_0 is normal, we have the estimate [7, p. 291]

$$dist(\Sigma(R), \Sigma(R_0)) \leq \|R_1\|.$$

Now if $|\Im z| > 1$, we have

$$|\sin(z)| \geq \frac{\exp(|\Im z|)}{4}, \quad |\cos(z)| \geq \frac{\exp(|\Im z|)}{4},$$

and for $|\Im z| \leq 1$, there is a $C > 0$ such that

$$|\sin(z)| \geq C\ dist(z, \{m\pi\}),\ |\cos(z)| \geq C\ dist\left(z, \left\{\left[m + \frac{1}{2}\right]\pi\right\}\right),\ m = 0, \pm 1, \pm 2, \ldots.$$

These estimates show that if $R(\lambda)$ has 0 as an eigenvalue for $|\lambda| \geq 1$, then $|\Im\omega| \leq C$, and for some k the diagonal entry $\rho_k(\lambda) \leq C|\omega|^{-2}$. Suppose $\rho_k \simeq \sin(\sqrt{\lambda - q_k})$ and consider λ satisfying $(n - 1/2)^2\pi^2 \leq |\lambda| < (n + 1/2)^2\pi^2$. If such a λ is an eigenvalue, then $\sqrt{\lambda - q_k} = n\pi + O(n^{-2})$, or $\lambda = n^2\pi^2 + q_k + O(n^{-1})$. The result is similar if $\rho_k \simeq \cos(\sqrt{\lambda - q_k})$. This shows that if λ is an eigenvalue, then $|\lambda - \lambda_{n,k}^0| = O(n^{-1})$ for some $\lambda_{n,k}^0$.

A perturbation argument [7, pp. 368–371] will show that for some C and $n \geq 1$, every disk $|\lambda - \lambda^0_{n,k}| \leq Cn^{-1}$ contains at least one eigenvalue. Consider the family of operators $L(t) = -D^2 + Q_0 + t[Q - Q_0]$ for $0 \leq t \leq 1$ whose domain is determined by the boundary conditions of (4a). Each of these operators on $\oplus_K L^2[0,1]$ has compact resolvent. The spectrum consists of the eigenvalues $\lambda^0_{n,k}$ when $t = 0$, and by the previous discussion, there is a family of open disks such as in the statement of the theorem whose complement is in the resolvent set for the entire family $L(t)$. If necessary, this sequence of disks may be modified so that any two disks $d_{n,k}$ are either the same or disjoint. Since the boundary of each disk is in the resolvent set for $0 \leq t \leq 1$, the corresponding family of spectral projections, which is continuous in t, has constant rank, and so each disk contains at least one eigenvalue.

The additional hypotheses provide improved estimates since they guarantee that $R_1 = O(\omega^{-3} e^{|\Im\omega|})$. This is clear from (3b) in case the off-diagonal entries of the matrices $Q(1)$, $Q(0)$, and $\int_0^1 Q(t) \int_0^t Q(s)\, ds\, dt$ vanish. In the cases of Dirichlet and Neumann boundary conditions, the trigonometric function in the third term of (3b) is $O(n^{-1})$ when $|\lambda - \lambda^0_{n,k}| = O(n^{-1})$. □

The conclusion of the previous theorem may be strengthened if $Q(x) = Q^*(x)$, since the operators $-D^2 + Q_0 + t[Q(x) - Q_0]$ associated to the eigenvalue problem (4a) form a self-adjoint holomorphic family on $\oplus_K L^2[0,1]$, each operator having compact resolvent. In this case [7, p. 392], the eigenvalues $\lambda^0_{n,k}$, counted with multiplicity, may be extended from $t = 0$ to $t = 1$, yielding the next result.

COROLLARY 4.2. *Suppose that Q_0 is a diagonal matrix, $Q(x) = Q^*(x)$, and $Q(x) \in C^2$. Then the eigenvalues of the problem* (4a)*, counted with multiplicity, may be indexed as $\lambda_{n,k}$ such that*

$$|\lambda_{n,k} - \lambda^0_{n,k}| \leq Cn^{-1}.$$

The supplementary conditions in Theorem 4.1 *will again improve these estimates to*

$$|\lambda_{n,k} - \lambda^0_{n,k}| \leq Cn^{-2}.$$

The next goal is to refine the eigenvalue estimates of Theorem 4.1 and Corollary 4.2. If $Q(x) \in C^J$, then a more precise description of $R(\lambda)$ may be obtained using the expansions of Lemmas 3.1 and 3.2. This expansion has the form

$$R(\lambda) = \sum_{j=0}^{J} \omega^{-j} [\alpha_j \cos(\omega) + \beta_j \sin(\omega)] + O(\omega^{-J-1} e^{|\Im\omega|}),$$

where α_j, β_j are constant $K \times K$ matrices. If Q_0 is a diagonal matrix with distinct diagonal entries, and $Q(x) \in C^J$, $J \geq 2$, then by Theorem 4.1 eigenvalues of (4a) with sufficiently large magnitude will be simple and may be indexed as $\lambda_{n,k}$ such that

$$|\lambda_{n,k} - \lambda^0_{n,k}| \leq Cn^{-1}.$$

The same remarks apply also for the large zeroes $\lambda^J_{n,k}$ of the function $\det R_J(\lambda)$, where

$$R_J(\lambda) = \sum_{j=0}^{J} \omega^{-j} [\alpha_j \cos(\omega) + \beta_j \sin(\omega)].$$

According to Theorem 4.1, the eigenvalues of (4a) are contained in disks δ_m of the form $\delta_m = \{|\lambda - m^2\pi^2| \le C\}$ or disks of the form $\{|\lambda - [m+1/2]^2\pi^2| \le C\}$. Since the cases are similar we will examine only the first case.

Consider the diagonal matrix $R_0(\lambda)$ with diagonal entries $\rho_k(\lambda)$ for λ inside a disk δ_m. Since $\sqrt{\lambda - q_k} = m\pi + O(m^{-1})$, the eigenvalues ρ_k of R_0 are $\pm 1 + O(m^{-2})$ or $O(m^{-1})$. For $\lambda \in \delta_m$ the matrices $R(\lambda)$ and $R_J(\lambda)$ are perturbations of $R_0(\lambda)$ satisfying

$$R(\lambda) - R_0(\lambda) = O(m^{-2}), \quad R_J(\lambda) - R_0(\lambda) = O(m^{-2}).$$

Since $R_0(\lambda)$ is normal, the spectra of the matrices $R(\lambda)$ and $R_J(\lambda)$ satisfy [7, p. 94]

(4b) $dist(\Sigma(R), \Sigma(R_0)) = O(m^{-2}), \quad dist(\Sigma(R_J), \Sigma(R_0)) = O(m^{-2}), \quad \lambda \in \delta_m.$

For m sufficiently large, and $\lambda \in \delta_m$, these eigenvalues may be partitioned into three groups, lying, respectively, in the disks $|\zeta| \le 1/4$ and $|\zeta \pm 1| \le 1/4$. For each of the matrices $R(\lambda)$, $R_0(\lambda)$, and $R_J(\lambda)$, the number of eigenvalues in each group, counted with algebraic multiplicity, is the same. For $l = 1, \dots, L$, denote by $\tau_l(\lambda)$, $\tau_l^0(\lambda)$, and $\tau_l^J(\lambda)$, respectively, the corresponding eigenvalues of $R(\lambda)$, $R_0(\lambda)$, and $R_J(\lambda)$ in the disk $|\zeta| \le 1/4$.

LEMMA 4.3. *Suppose that Q_0 has distinct diagonal entries, and $Q(x) \in C^J$, $J \ge 2$. For $\lambda \in \delta_m$ with m sufficiently large, the eigenvalues $\tau_l(\lambda)$, $\tau_l^0(\lambda)$, and $\tau_l^J(\lambda)$ are simple and analytic in λ. In addition*

$$\tau_l(\lambda)' = \frac{\pm 1}{2m\pi} + O(m^{-2})$$

in the disks $\tilde{\delta}_m$ centered at $m^2\pi^2$ with half the radius of δ_m.

Proof. By definition of R_0

$$\pm\tau_l^0(\lambda) = (1 + \omega^{-2}[2^{-3}q_l^2 + c_l])\sin(\sqrt{\lambda - q_l}).$$

A Taylor expansion of the sine function near $\lambda = m^2\pi^2$ shows that for m sufficiently large and $\lambda \in \delta_m$

$$|\tau_k^0 - \tau_l^0| \ge \frac{|q_k - q_l|}{4m}.$$

By (4b) the same type of estimate holds for $\tau_l(\lambda)$ and $\tau_l^J(\lambda)$ as well. The analyticity now follows from the analyticity of the matrices R, R_0, and R_J.

The derivative estimate follows from the Cauchy integral representation

$$\tau_l(\lambda)' - \tau_l^0(\lambda)' = \frac{1}{2\pi i}\int_{\partial\delta_m} \frac{\tau_l(z) - \tau_l^0(z)}{(z-\lambda)^2}\, dz.$$

Since

$$|\tau_l(\lambda) - \tau_l^0(\lambda)| = O(m^{-2})$$

for $\lambda \in \delta_m$, it follows that

$$|\tau_l(\lambda)' - \tau_l^0(\lambda)'| = O(m^{-2})$$

inside the disks $\tilde{\delta}_m$. The desired estimate follows from the elementary calculation $\tau_l^0(\lambda)' = \pm 1/(2m\pi) + O(m^{-3})$. □

THEOREM 4.4. *Suppose that Q_0 is a diagonal matrix with distinct diagonal entries, and $Q(x) \in C^J$, $J \geq 2$. For n sufficiently large the eigenvalues $\lambda_{n,k}$ of* (4a) *satisfy*

$$|\lambda_{n,k} - \lambda_{n,k}^J| \leq Cn^{-J}.$$

Proof. Our main goal is to show that the resolvent $[R(\lambda) - \zeta]^{-1}$ satisfies an estimate

$$\|[R(\lambda) - \zeta]^{-1}\| \leq \frac{C}{dist(\zeta, \Sigma(R))}$$

for $|\zeta| \leq 1/4$, the estimate holding uniformly for $\lambda \in \tilde{\delta}_m$ for m large (and in corresponding disks centered at $[m + 1/2]^2\pi^2$).

Define the projections

$$P_0 = \frac{-1}{2\pi i} \int_{|\zeta|=1/4} [R(\lambda) - \zeta]^{-1}\, d\zeta, \quad P_{\pm 1} = \frac{-1}{2\pi i} \int_{|\zeta \pm 1|=1/4} [R(\lambda) - \zeta]^{-1}\, d\zeta.$$

These projections may be used to decompose the operator R [7, pp. 40–43]. Since $R(\lambda) - R_0(\lambda) = O(m^{-2})$ and $\|[R_0(\lambda) - \zeta]^{-1}\| = 1/dist(\zeta, \Sigma(R_0))$, the expression

$$[R(\lambda) - \zeta]^{-1} = [R_0(\lambda) - \zeta]^{-1}[I + (R - R_0)[R_0 - \zeta]^{-1}]^{-1} \tag{4c}$$

shows that the projections $P_0, P_{\pm 1}$ converge to the corresponding orthogonal projections $\widetilde{P_0}, \widetilde{P_{\pm 1}}$ for R_0. The convergence is uniform for large m and $\lambda \in \delta_m$.

Decompose $X \in \mathbf{C}^K$ as $X = X_0 + X_1 + X_{-1}$ where X_i is in the range of P_i. Since $R(\lambda)$ is bounded in the disks δ_m, and $P_i(\lambda) \to \widetilde{P_i(\lambda)}$, we have, for $i = \pm 1$,

$$\begin{aligned} \|(R - \zeta)X_i\| &\geq \|(R - \zeta)\widetilde{P}_i X_i\| - \|(R - \zeta)(P_i - \widetilde{P}_i)X_i\| \\ &\to |i - \zeta| \|\widetilde{P}_i X_i\|, \quad |\zeta| \leq \frac{1}{4}. \end{aligned}$$

The convergence $P_i(\lambda) \to \widetilde{P_i(\lambda)}$ then implies

$$\|(R - \zeta)X_i\| \geq \frac{1}{4}\|X_i\|$$

for m large, or

$$\|[R(\lambda) - \zeta]^{-1} P_i\| \leq 4, \quad \lambda \in \delta_m, \quad |\zeta| \leq \frac{1}{4}.$$

We next consider $\|[R(\lambda) - \zeta]^{-1} P_0\|$, for $\lambda \in \delta_m$, and $|\zeta| \leq 1/4$. The proof of Lemma 4.3 shows that for m large and $\lambda \in \delta_m$, the eigenvalues τ_l^0 of R_0 are distinct and $|\tau_k^0 - \tau_l^0| \geq C/m$. Let γ_l be a collection of circular contours centered at τ_l^0 with radius $C/(4m)$. Define projections

$$\mathcal{P}_l = \frac{-1}{2\pi i} \int_{\gamma_l} [R(\lambda) - \zeta]^{-1}\, d\zeta, \quad \widetilde{\mathcal{P}_l} = \frac{-1}{2\pi i} \int_{\gamma_l} [R_0(\lambda) - \zeta]^{-1}\, d\zeta.$$

Since $\|R(\lambda) - R_0(\lambda)\| = O(m^{-2})$ the expression (4c) leads to the estimate

$$\|\mathcal{P}_l - \widetilde{\mathcal{P}_l}\| = \left\| \frac{-1}{2\pi i} \int_{\gamma_l} [R(\lambda) - \zeta]^{-1} - [R_0(\lambda) - \zeta]^{-1} \, d\zeta \right\| = O(m^{-1}), \quad \lambda \in \delta_m.$$

Thus, the part of the operator $R(\lambda)$ on the L-dimensional range of P_0 has L distinct eigenvalues and may be diagonalized with a matrix S whose columns are eigenvectors of $R(\lambda)$. The comparison of $\mathcal{P}_l$ with $\widetilde{\mathcal{P}_l}$ shows that these eigenvectors may be chosen to converge to standard basis vectors, so that $\|S(\lambda)\|$ and $\|S^{-1}(\lambda)\|$ converge to 1 as $\lambda \to \infty$. This easily gives the estimate

$$\|[R(\lambda) - \zeta]^{-1} P_0\| \leq \frac{C}{dist(\zeta, \Sigma(R(\lambda))}, \quad |\zeta| \leq \frac{1}{4}.$$

Putting the pieces together we find

$$\begin{aligned} \|[R(\lambda) - \zeta]^{-1}\| &= \|[R(\lambda) - \zeta]^{-1}[P_0 + P_1 + P_{-1}\| \\ &\leq \|[R(\lambda) - \zeta]^{-1} P_0\| + \|[R(\lambda) - \zeta]^{-1} P_1\| + \|[R(\lambda) - \zeta]^{-1} P_{-1}\| \\ &\leq \frac{C}{dist(\zeta, \Sigma(R(\lambda))} \end{aligned}$$

as desired.

To conclude the proof, recall that $R(\lambda) - R_J(\lambda) = O(m^{-J-1})$. Similar to the argument about perturbations of the normal matrix R_0, we find that $|\tau_l - \tau_l^J| = O(m^{-J-1})$. Now Lemma 4.3 states that $\tau_l(\lambda)' = \pm 1/(2m\pi) + O(m^{-2})$ for $\lambda \in \tilde{\delta_m}$. For some l, $\tau_l(\lambda_{n,k}) = 0$, which implies $\tau_l^J(\lambda_{n,k}) = O(m^{-J-1})$ or

$$|\lambda_{n,k} - \lambda_{n,k}^J| = O(m^{-J}). \qquad \square$$

The eigenvalue estimate of Corollary 4.2 for the Dirichlet boundary conditions $Y(0) = 0 = Y(1)$ was conjectured in [11] and served as the basis for development of a trace formula. We develop an alternative approach to trace formulas of any order based on Lemma 2.1 and some observations about the entire function $\det(\mathcal{T}(\lambda))$.

Suppose first that Q_0 is diagonal and $Q(x) \in C^2$. Since the determinant of R has the form

$$\det(R) = \sum_{\sigma} \prod_{k=1}^{K} (-1)^{sgn(\sigma)} r_{k,\sigma(k)},$$

the sum taken over permutations σ of $1, \ldots, K$, each summand is either the product of all the diagonal entries or has two off-diagonal factors. Each off-diagonal entry of R is $O(\omega^{-2} e^{|\Im\omega|})$, so that any product in the determinant with two off-diagonal factors is $O(\omega^{-4} e^{K|\Im\omega|})$.

This analysis may be improved when $Q(x) \in C^J$, and the expansions of Lemmas 3.1 and 3.2 refine the estimates (3b). Recall that $p_j(\omega)$ is a trigonometric polynomial of degree at most K if

$$p_j(\omega) = \sum_{k=-K}^{K} c_k e^{ik\omega}.$$

The next result summarizes the nature of $\det(\mathcal{T}(\lambda))$.

THEOREM 4.5. *The function* $\det(\mathcal{T}(\lambda))$ *is an entire function of order* $1/2$. *The zeroes of* $\det(\mathcal{T}(\lambda))$ *are precisely the eigenvalues of* (4a), *and their orders agree with the algebraic multiplicity of the eigenvalue. If* $Q(x) \in C^J$, *then for some integer* L

$$\det(\mathcal{T}(\lambda)) = \omega^L \Big[\sum_{j=0}^{J} \omega^{-j} p_j(\omega) + O(\omega^{-J-1} e^{K|\Im\omega|}) \Big],$$

where the functions p_j *are trigonometric polynomials of degree at most* K *whose coefficients are determined by the coefficients* $A_j(1), \ldots, D_j(1)$, *for* $j \leq J$ *of Lemmas* 3.1 *and* 3.2. *In particular if* $J \geq 2$, *then*

$$\det(\mathcal{T}(\lambda)) = \omega^L \Big[\prod_{k=1}^{K} \rho_k + O(\omega^{-3} e^{K|\Im\omega|}) \Big].$$

The only claim in Theorem 4.5 that requires further comment is that the order of the zeroes of $\det(\mathcal{T}(\lambda))$ agrees with the algebraic multiplicity of the eigenvalues of (4a). We sketch the argument. The claim is easy to check directly if $Q(x)$ is diagonal and all eigenvalues are simple. In the general case connect $Q(x)$ to a diagonal coefficient $Q_1(x)$ with all eigenvalues simple by an analytic path $Q_1(x) + z[Q(x) - Q_1(x)]$. The eigenvalues (depending now on z) for (4a) may be partitioned into finite systems, and for each system [7, p. 370] the eigenvalues are simple except for a finite set of z with $|z| < 2$. As long as all the eigenvalues in a finite system remain simple, the result holds by analytic continuation, and the exceptional points are handled by continuity.

We are now prepared to compute the traces appearing in Lemma 2.1.

THEOREM 4.6. *Suppose that* $Q(x), \widetilde{Q}(x) \in C^J$, $J \geq 2$, *and that the matrices* $R_J(\lambda)$ *and* $\widetilde{R}_J(\lambda)$ *agree. The matrix* Q_0 *is assumed to have distinct diagonal entries.*

The respective eigenvalues μ_n *and* ν_n *of the problems* (4a) *may be ordered such that*

$$\sum_n |\mu_n^l - \nu_n^l| < \infty, \quad l = 1, \ldots, \left\lfloor \frac{J}{2} \right\rfloor.$$

Moreover, the traces

$$\sum_n \mu_n^l - \nu_n^l$$

may be expressed as polynomials in the coefficients of the trigonometric polynomials p_j *of Theorem* 4.5.

Proof. The arguments of Theorem 4.4 show that there is a $C_1 > 0$ such that the eigenvalues μ_n and ν_n with magnitude greater than C_1 may be paired by choosing the closest member from the other sequence. In addition C_1 may be chosen so that the circle $|\lambda| = C_1$ is in the resolvent set for all the operators $-D^2 + \widetilde{Q}(x) + t[Q(x) - \widetilde{Q}(x)]$ for $0 \leq t \leq 1$. It follows that the algebraic multiplicity for the system of eigenvalues inside the circle is independent of t, and these eigenvalues μ_n and ν_n, represented with algebraic multiplicity, may be paired arbitrarily.

With this indexing scheme Theorem 4.4 shows that

$$|\mu_n - \nu_n| = O(n^{-J}).$$

Since $|\mu_n| \le Cn^2$ we have

$$\sum_n |\mu_n - \nu_n|[|\mu_n|^{j-1-\epsilon} + |\mu_n|^{j-1-\epsilon}] < \infty,$$

as long as $j \le (J+2)/2$ and $\epsilon > 1/2$. Lemma 2.1 now applies to give the existence of the traces.

By Theorem 4.5 and the Hadamard product theorem [1], the functions $\det(\mathcal{T}(\lambda))$ and $\det(\widetilde{\mathcal{T}}(\lambda))$ play the role of the functions f and g of Lemma 2.1. What remains, then, is to compute the asymptotics for

$$\log\Big(\frac{\det(\mathcal{T}(\lambda))}{\det(\widetilde{\mathcal{T}}(\lambda))}\Big)$$

as $\lambda \to -\infty$ along the real axis. Divide $\det(\mathcal{T}(\lambda))$ and $\det(\widetilde{\mathcal{T}}(\lambda))$ by $\omega^L \exp(K|\Im\omega|)$. By Theorem 4.5 the expression

$$\frac{\det(\mathcal{T}(\lambda))}{\omega^L \exp(-iK\Im\omega)}$$

and the corresponding expression for $\widetilde{\mathcal{T}}$ are a sum of a polynomial in ω^{-1} of degree J and terms that decay like $O(\omega^{-J-1})$ as $\lambda \to -\infty$. In addition, the order zero coefficients are not zero. The computations are completed using a power series expansion for $\log(1+z)$. □

These techniques may be used to resolve a question [11] about the computation of traces for systems of the form (4a) with the Dirichlet boundary conditions $Y_i(0) = 0 = Y_i(1)$. In this case it is sufficient to assume that Q_0 is similar to a diagonal matrix since the application of the similarity transformation to (4a) will leave the boundary conditions fixed.

THEOREM 4.7. *Suppose that $Q \in C^2$ is real symmetric and the boundary conditions $Y_i(0) = 0 = Y_i(1)$ are used for the problem* (4a). *Let $\lambda_{n,k}$ be the eigenvalues with coefficient $Q(x)$ and $\lambda^0_{n,k}$ be the eigenvalues with coefficient Q_0. Then*

$$\sum_n \sum_k (\lambda_{n,k} - \lambda^0_{n,k}) = \operatorname{tr} \frac{Q(1) + Q(0) - 2Q_0}{4}.$$

Proof. Consider the two functions $\mathcal{T}(\lambda)$ and $\mathcal{T}_0(\lambda)$ associated, respectively, with the coefficients $Q(x)$ and its integral Q_0. From (3b) the constants c_k in the diagonal entries ρ_k of $R_0(\lambda)$ are the diagonal entries α_k of

$$2^{-2}\Big[Q(1) + Q(0) - \int_0^1 Q(t) \int_0^t Q(s)\, ds\, dt\Big],$$

for $Q(x)$, and they are the diagonal entries $\beta_k = 2^{-1}q_k - 2^{-3}q_k^2$ of $2^{-1}Q_0 - 2^{-3}Q_0^2$ for the diagonal matrix Q_0. Theorem 4.5 shows that

$$\frac{\det(\mathcal{T}(\lambda))}{\det(\mathcal{T}_0(\lambda))} = \frac{\prod_k (1 + \omega^{-2}[2^{-3}q_k^2 - \alpha_k]) + O(\omega^{-3})}{\prod_k (1 + \omega^{-2}[2^{-3}q_k^2 - \beta_k]) + O(\omega^{-3})},$$

and a Taylor expansion gives

$$\log\Big(\frac{\det(\mathcal{T}(\lambda))}{\det(\mathcal{T}_0(\lambda))}\Big) = \lambda^{-1} \sum_k (\beta_k - \alpha_k) + O(\omega^{-3}).$$

By Corollary 4.2, the Dirichlet boundary conditions imply that $|\lambda_{n,k} - \lambda^0_{n,k}| = O(n^{-2})$. Lemma 2.1 applies with $1/2 < \epsilon < 1$ to give

$$\sum_n \sum_k (\lambda_{n,k} - \lambda^0_{n,k}) = \sum_k (\beta_k - \alpha_k).$$

Notice that if A and B are symmetric matrices, then

$$(AB)_{ii} = \sum_j a_{ij} b_{ji} = \sum_j a_{ji} b_{ij} = \sum_j b_{ij} a_{ji} = (BA)_{ii}.$$

Thus, if $Q(t)$ is symmetric, the diagonal entries of $\int_0^1 Q(t) \int_0^t Q(s)\ ds\ dt$ will agree with those of $\int_0^1 \int_0^t Q(s)\ dsQ(t)\ dt$. Differentiation gives

$$\partial_x \left(\int_0^x Q(t)\ dt \right)^2 = Q(x) \int_0^x Q(t)\ dt + \int_0^x Q(t)\ dtQ(x),$$

and integration shows that the diagonal entries of $\int_0^1 Q(t) \int_0^t Q(s)\ ds\ dt$ are $2^{-1} q_k^2$. These common contributions drop out of the differences $\alpha_k - \beta_k$, showing that

$$\sum_n \sum_k (\lambda_{n,k} - \lambda^0_{n,k}) = \mathrm{tr} \frac{Q(1) + Q(0) - 2Q_0}{4}. \qquad \square$$

REFERENCES

[1] L. Ahlfors, *Complex Analysis*, McGraw-Hill, New York, 1966.

[2] L.A. Dikii, *Trace formulas for Sturm-Liouville differential operators*, Amer. Math. Soc. Transl., 18 (1958), pp. 81–115.

[3] C. Fulton and S. Pruess, *Eigenvalue and eigenfunction asymptotics for regular Sturm-Liouville problems*, J. Math. Anal. Appl., 188 (1994), pp. 297–340.

[4] J. Garnett and E. Trubowitz, *Gaps and bands of one-dimensional periodic Schrödinger operators*, Comment. Math. Helvetici, 59 (1984), pp. 258–321.

[5] F. Gesztesy and H. Holden, *On trace formulas for Schrödinger-type operators*, Multiparticle Quantum Scattering with Aaplications to Nuclear, Atomic and Molecular Physics, IMA Vol. Math. Appl. 89, Springer, New York, 1997, pp.121–145.

[6] F. Gesztesy and B. Simon, *The Xi function*, Acta Math., 176 (1996), pp. 49–71.

[7] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1995.

[8] W. Magnus and S. Winkler, *Hill's Equation*, Dover, New York, 1979.

[9] V. Marchenko, *Sturm-Liouville Operators and Applications*, Birkhäuser, Basel, 1986.

[10] H.P. McKean and P. van Moerbeke, *The spectrum of Hill's equation*, Invent. Math., 30 (1975), pp. 217–274.

[11] V.G. Papanicolaou, *Trace formulas and the behaviour of large eigenvalues*, SIAM J. Math. Anal., 26 (1995), pp. 218–237.

[12] E. Trubowitz, *The inverse problem for periodic potentials*, Comm. Pure Appl. Math., 30 (1977), pp. 321–337.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 963–984

ATTRACTORS OF SOME REACTION DIFFUSION PROBLEMS*

JACK K. HALE† AND JOSÉ DOMINGO SALAZAR GONZÁLEZ‡

Abstract. In this paper we study the existence and stability properties of certain solutions of a semilinear parabolic equation with Robin boundary conditions. We are actually interested in solutions that exhibit both boundary and internal layers. We give an extension of the Sturm–Liouville theory to treat this problem and compute the number of stable solutions. We also completely determine the attractor for a few examples. Finally, we show that our results are robust and that, in particular, the structure of these attractors persist under small perturbations.

Key words. reaction diffusion, singular perturbation, attractors, time maps

AMS subject classifications. 35K57, 35B25, 35B35

PII. S0036141097327641

1. Introduction.

If $u(x,t)$ is a smooth function, we let $\dot{u} = \partial u/\partial t$, $u' = \partial u/\partial x$, $u'' = \partial^2 u/\partial x^2$. We are interested in the study of the flow defined by the scalar reaction diffusion equation

$$\dot{u} = \epsilon^2 u'' + f(x,u), \qquad 0 < x < 1, \quad t \geq 0, \quad \epsilon > 0, \tag{1}$$

with the boundary conditions

$$\begin{cases} \alpha_0 u(0,t) - (1-\alpha_0)u'(0,t) = \beta_0, \\ \alpha_1 u(1,t) + (1-\alpha_1)u'(1,t) = \beta_1, \end{cases} \tag{2}$$

where $0 \leq \alpha_0, \alpha_1 \leq 1$ and β_0, β_1 are real constants.

The function f is assumed to be smooth in u and we consider the solutions of (1) with initial value in $H^1(0,1)$. We suppose also that f satisfies a dissipative condition for large u which will ensure that there is a compact global attractor. This will be the case, for example, if f and u have opposite signs for large values of $|u|$. A typical example and the one that will be the center of our attention in this paper is the case where f is a cubic polynomial in u and is given explicitly by

$$f(x,u) = u(1-u)(u-c(x)), \qquad 0 < c(x) < 1. \tag{3}$$

Zelenyak [14] proved that the ω-limit set of each solution of (1) is an equilibrium point; that is, a solution of the equation

$$\epsilon^2 u'' + f(x,u) = 0, \qquad 0 < x < 1, \quad t \geq 0, \quad \epsilon > 0,$$

with the boundary conditions (2).

For homogeneous boundary conditions (2) ($\beta_0 = \beta_1 = 0$) and $f(x,u)$ independent of x, it is a consequence of results of Yanagida [13] that any nonconstant equilibrium

*Received by the editors September 22, 1997; accepted for publication (in revised form) July 23, 1998; published electronically July 7, 1999.

http://www.siam.org/journals/sima/30-5/32764.html

†Georgia Institute of Technology, School of Mathematics, Atlanta, GA 30332-0160 (hale@math.gatech.edu). The research of this author was supported in part by ARO grant DAAG-55-98-1-0364 and NSF grant DMS-9704853.

‡University of Oxford, Mathematical Institute, 24–29 St. Giles', Oxford OX1 3LB, UK (salazar@maths.ox.ac.uk). The research of this author was supported in part by NSF grant DMS-9306265.

solution u of (1) is unstable if it has the property that there are two values in $[0,1]$ for which $u'=0$.

In particular, if we take homogeneous Neumann boundary conditions ($\alpha_0=\alpha_1=0$), we deduce that any nonconstant equilibrium solution is unstable. Therefore, the stable solutions of (1) with homogeneous Neumann boundary conditions are the zeros of $f(u)$ which are stable as solutions of the ODE $\dot{u}=f(u)$. In particular, for the cubic in (3) with $c(x)=c_0\in(0,1)$ with c_0 constant, the only stable solutions are the constant functions 0 and 1.

If we consider (1) and (3) with homogeneous Neumann boundary conditions and allow $c(x)$ to depend upon x but still belong to the open interval $(0,1)$, then the problem becomes much more complicated, and there are situations where it is possible to obtain stable nonconstant equilibrium solutions. In fact, suppose that $c\in C^1([0,1])$ with $c\neq 0,\ c'\neq 0$ at $x=0,\ x=1$ and, if $c(x)=1/2$, then $c'(x)\neq 0$. If $c(x)=1/2$ at M points in the interval $(0,1)$, then there is an $\epsilon_0>0$ such that, for every $\epsilon\in(0,\epsilon_0)$; there is exactly the Mth Fibonacci number of stable solutions (Angenent, Mallet-Paret, and Peletier [1]). A stable nonconstant solution u has the property that, if it takes the value $1/2$ near some point x_0 where $c(x_0)=1/2$, then it develops a sharp transition layer at x_0 as $\epsilon\to 0$ and $c'(x_0)u'(x_0)<0$. There are equilibrium solutions u_0 for which there is a sharp transition layer as above at x_0 and $c'(x_0)u'(x_0)>0$ (Hale and Sakamoto [6]). Kwapisz [8] has obtained some extensions to the case where the graph of c and the graph of the constant function $1/2$ are not transversal. Kurland [7] proved that there could be highly oscillatory solutions (automatically unstable) at the points where $c=1/2$.

If the function $c(x)$ is a step function, Rocha [12] has obtained the same result as Angenent, Mallet-Paret, and Peletier [1] with the number M being the number of times that the function c jumps across $1/2$. His method also allows one to obtain all of the equilibrium solutions since it is based upon phase plane methods. In this case, under certain conditions on the jumps, the number of equilibrium solutions is bounded independently of ϵ in contrast to the situation for smooth functions c.

In this paper, we consider the case with c a step function and the general boundary conditions (2). The objective is to understand how the index of an equilibrium point depends upon the boundary conditions. As we will see, there are limitations on the number of stable equilibrium solutions as well as the total number of equilibrium solutions. We follow the methods of Rocha [11, 12], making the appropriate generalizations to the more general boundary conditions. The main result is contained in Theorem 3.3. We also give numerical results which indicate the method used to obtain the solutions as well as yield the permutation matrix of Fiedler and Rocha [4] which gives the manner in which the equilibrium points are connected by heteroclinic orbits. Fiedler and Rocha [3] also have shown that this connection matrix characterizes the topological properties of the flow on the compact global attractor; that is, if c_1, c_2 have the same connection matrix, then the flows on the compact global attractors are topologically equivalent.

Although our cubic nonlinearity (3) depends on x, we have that for every value of x, $f(x,0)=f(x,1)=0$. Also note that since

$$f_u(x,u)=-3u^2+2(c(x)+1)u-c(x),$$

the derivatives with respect to u of f at $u=0$ and at $u=1$ are

$$f_u(x,0)=-c(x)\qquad\text{and}\qquad f_u(x,1)=c(x)-1,$$

so they are both negatives due to (3).

As we said before, this problem has a gradient structure that guarantees that the attractor consists of equilibrium solutions and their unstable manifolds. Also, it is known that the ω-limit set of any bounded solution is a singleton [9, 14]. Thus, in order to gain information about the dynamics of this problem, we are going to concentrate our attention on the stationary solutions. They verify the following boundary value problem for $u = u(x)$ ($0 \le \alpha_0, \alpha_1 \le 1$; $\beta_0, \beta_1 \in \mathbb{R}$):

$$\begin{cases} \epsilon^2 u'' + f(x,u) = 0, & 0 < x < 1, \quad \epsilon > 0, \\ \alpha_0 u(0) - (1-\alpha_0) u_x(0) = \beta_0, \\ \alpha_1 u(1) + (1-\alpha_1) u_x(1) = \beta_1. \end{cases} \tag{4}$$

2. Sturm–Liouville properties. Following Rocha [11, 12], we are going to characterize the existence and hyperbolic nature of the equilibria of (1) in the case of homogeneous conditions, by defining some appropriate angles in the phase planes corresponding to (4) and its variational equation as is done is the classical Sturm–Liouville theory. For the nonhomogeneous case, we will be able to define only the angle corresponding to the variational equation. Since the results are identical, we will write $a(x)$ instead of ϵ^2, where $a : [0,1] \to \mathbb{R}$ with $a(x) > 0$.

Thus, in most of this section we are going to be concerned with the study of the equilibria of the problem

$$\begin{cases} u_t = (a(x)u_x)_x + f(x,u), & 0 < x < 1, \\ \alpha_0 u(0) - (1-\alpha_0) u_x(0) = 0, \\ \alpha_1 u(1) + (1-\alpha_1) u_x(1) = 0. \end{cases} \tag{5}$$

Let us represent by $\mathcal{E}$ the set of stationary solutions of (5). Thus, if $u = u(x) \in \mathcal{E}$, then u will have to satisfy

$$\begin{cases} (a(x)u')' + f(x,u) = 0, & 0 < x < 1, \\ \alpha_0 u(0) - (1-\alpha_0) u_x(0) = 0, \\ \alpha_1 u(1) + (1-\alpha_1) u_x(1) = 0 \end{cases} \tag{6}$$

for $0 \le \alpha_0, \alpha_1 < 1$. The cases in which α_0 or α_1 are equal to 1 will be considered in a remark at the end of this section.

In order to study this boundary value problem, we are going to set up a *shooting method* and for that we will work with the equivalent first-order system of equations

$$\begin{cases} u_x = v/a(x), & v_x = -f(x,u); \\ u(0) = u_0, & v(0) = \alpha_0 u_0/(1-\alpha_0). \end{cases} \tag{7}$$

The maximum principle allows us to conclude that the equilibria are in the interval $[0,1]$. Therefore, we need only consider $u_0 \in [0,1]$. Changing to polar coordinates $u := p\cos q$, $v := -p\sin q$, the angle $q := q(x,u_0)$ will satisfy

$$\begin{cases} q_x = \sin^2 q/a(x) + (1-u)(u-c(x))\cos^2 q, & 0 < x < 1, \\ q(0,u_0) = q_0, \end{cases} \tag{8}$$

where $-\pi/2 < q_0 < 0$ is the corresponding initial angle that verifies $\tan q_0 = -\alpha_0/(1-\alpha_0)$.

If we define $\sigma(u_0) := q(1,u_0)$, then $\sigma : [0,1] \to (-\pi/2, +\infty)$ and we arrive at the following proposition.

PROPOSITION 2.1. *The set of nontrivial equilibria of* (5) *is in one-to-one correspondence with the set*

$$\{u_0 : \sigma(u_0) = q_1 + k\pi, \quad k \in \mathbb{N} \cup \{0\}\},$$

where $\tan q_1 = \alpha_1/(1-\alpha_1)$.

Proof. This is a simple consequence of the flow direction of (8) and the general Sturm–Liouville theory. □

We can apply a similar analysis to the variational equation

$$\begin{cases} \eta_x = \mu/a(x), & \mu_x = -f_u(x,u)\eta; \\ \eta(0) = 1, & \mu(0) = \alpha_0/(1-\alpha_0), \end{cases} \tag{9}$$

where $\eta := du/du_0$ and $\mu := a(x)\eta_x$. If we make the change to the polar coordinates $\eta := \psi\cos\phi$, $\mu := -\psi\sin\phi$, then the angle ϕ will verify

$$\begin{cases} \phi_x = \sin^2\phi/a(x) + f_u(x,u)\cos^2\phi, & 0 < x < 1, \\ \phi(0,u_0) = q_0. \end{cases} \tag{10}$$

Now we define the function $\theta(u_0) := \phi(1,u_0)$, so $\theta : [0,1] \to (-\pi/2, +\infty)$. This function will allow us to determine the stability properties of the equilibria of (5) as stated in the following theorem.

THEOREM 2.2. *An equilibrium point* $u = u(\cdot,u_0)$ *of* (5) *is hyperbolic if and only if* $\theta(u_0) \neq q_1 + k\pi$ *for any* $k \in \mathbb{N} \cup \{0\}$. *Moreover, if* $W^u(u)$ *denotes the unstable manifold of* $u = u(\cdot,u_0) \in \mathcal{E}$ *and* u *is hyperbolic, then*

$$\dim W^u(u) = 1 + \left[\frac{\theta(u_0) - q_1}{\pi}\right],$$

where $[\cdot]$ *represents the integer part. Consequently,* $u = u(\cdot,u_0) \in \mathcal{E}$ *is hyperbolic and asymptotically stable if and only if* $\theta(u_0) < q_1$.

Proof. These results are a consequence of comparing the variational problem (9) with the eigenvalue problem for $\lambda = 0$ around an equilibrium solution $u = u(x,u_0)$ $(u \longmapsto u + w)$

$$\begin{cases} (a(x)w')' + f_u(x,u)w = \lambda w, & 0 < x < 1, \\ \alpha_0 w(0) - (1-\alpha_0)w_x(0) = 0, \\ \alpha_1 w(1) + (1-\alpha_1)w_x(1) = 0; \end{cases} \tag{11}$$

written as a first-order system with $\nu = a(x)w_x$,

$$\begin{cases} w_x = \nu/a(x), & \nu_x = (\lambda - f_u(x,u))w; \\ w(0) = w_0, & \nu(0) = \alpha_0 w_0/(1-\alpha_0), \end{cases}$$

and then changed to polar coordinates $(w := z\cos\zeta,\ \nu := -z\sin\zeta)$ to obtain

$$\begin{cases} \zeta_x = \sin^2\zeta/a(x) + (\lambda - f_u(x,u))\cos^2\zeta, & 0 < x < 1, \\ \zeta(0,\lambda) = q_0, \quad \zeta(1,\lambda) = q_1 + n\pi. \end{cases}$$

Thus for $\lambda = 0$ the eigenvalue angle equation is identical to (10).

In order to prove the rest of the theorem, we have to recall from the Sturm–Liouville theory (see [2]) that $\zeta(1,\lambda)$ is a strictly decreasing function of λ and that

the eigenvalues are ordered by the number of zeros of their corresponding eigenfunctions. ∎

Additionally, we can obtain the following characterizations of the nonhyperbolic equilibria in terms of the end-point angles σ and θ.

LEMMA 2.3. *Let $u = u(\cdot, u_0)$ be an equilibrium point not identically zero. Then $\theta(u_0) = q_1 + k\pi$ if and only if $\sigma'(u_0) = 0$.*

LEMMA 2.4. *A critical point u_0 of σ (that is, $\sigma'(u_0) = 0$) is nondegenerate (so $\sigma''(u_0) \neq 0$) if and only if $\theta'(u_0) \neq 0$.*

Proof. If we make the following changes to polar coordinates: for $u = u(x, u_0)$, $v = a(x)u_x$, $\eta = du/du_0$, and $\mu = a(x)\eta$, we have

$$\begin{cases} u = p\cos q, \\ v = -p\sin q, \end{cases} \qquad \begin{cases} \eta = \psi\cos\phi, \\ \mu = -\psi\sin\phi. \end{cases}$$

Differentiating the first two equations with respect to u_0 and equating them with respect to the second ones, we deduce that

$$\begin{aligned} \frac{\partial p}{\partial u_0}\cos q - p\sin q\frac{\partial q}{\partial u_0} &= \psi\cos\phi, \\ \frac{\partial p}{\partial u_0}\sin q + p\cos q\frac{\partial q}{\partial u_0} &= \psi\sin\phi, \end{aligned}$$

which, eliminating $\partial p/\partial u_0$ in these equations and setting $x = 1$, leads us to

$$r\sigma' = \rho\sin(\theta - \sigma), \tag{12}$$

where $r(u_0) := p(1, u_0)$ and $\rho(u_0) := \psi(1, u_0)$. Because $(0, 0)$ is an equilibrium point of (7) and (9), r and ρ cannot be zero if $u(x, u_0)$ is not identically zero. Thus if $u(x, u_0)$ is an equilibrium point then $\sigma(u_0) = q_1 + k\pi$; if we assume $\theta(u_0) = q_1 + k\pi$ also, then $\sin(\theta - \sigma) = 0$ and vice versa. From here we can obtain the conclusion of Lemma 2.3.

If now we differentiate (12) with respect to u_0, we will get

$$r\sigma'' = \pm\rho\theta'$$

for any critical point of σ. From here, the conclusion of Lemma 2.4 is clear. ∎

Let us now define the *lifted manifold* $\mathcal{M}$ to be the solution manifold $\mathcal{M} = \{(x, u, v) : u = u(x, u_0), v = v(x, u_0) \text{ for } u_0 \in [0, 1]\}$. Let us also define L_y to be the section curve of $\mathcal{M}$ at $x = y$. Then the function σ is the angle that a point of the curve L_1 makes with the u-axis and θ corresponds to the angle of the tangent at that point. See Figure 1.

We can obtain similar results to the previous ones if we consider the backward shooting method

$$\begin{cases} \bar{u}_x = \bar{v}/a(x), & \bar{v}_x = -f(x, u); \\ \bar{u}(1) = u_0, & \bar{v}(1) = -\alpha_1 u_0/(1 - \alpha_1). \end{cases} \tag{13}$$

Then if we define $\bar{\mathcal{M}}$ and $\bar{L}_y$ relative to problem (13), we can prove the following proposition.

PROPOSITION 2.5. *There is a one-to-one correspondence between the set of equilibria $\mathcal{E}$ and the set $L_y \cap \bar{L}_y$ for any $y \in [0, 1]$.*

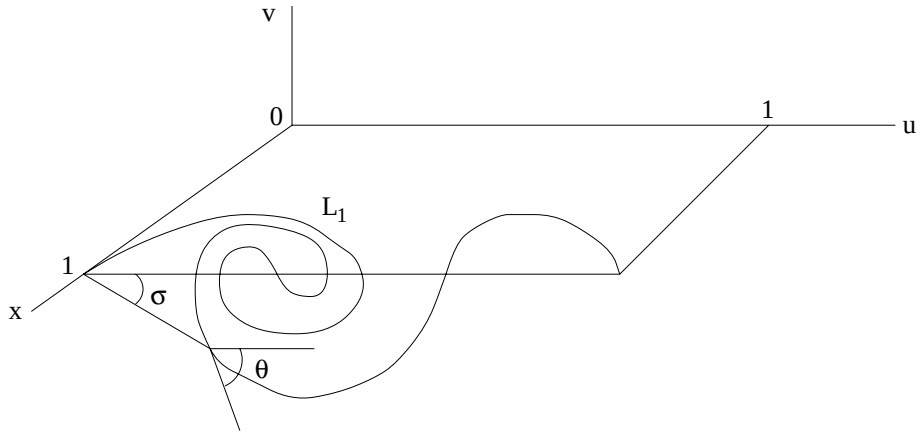


FIG. 1. *Time map* L_1.

Finally, we obtain stability information about a particular equilibrium $u = u(\cdot, u_0)$ by looking at the intersection $L_y \cap \bar{L}_y$. Actually, let $\bar{\phi} = \bar{\phi}(x, u_0)$ be the angle corresponding to the backward version of the variational problem (10) which has the initial value $\bar{\phi}(1, u_0) = q_1$. If we now define $\Omega(y, u_0) = \phi(y, u_0) - \bar{\phi}(y, u_0)$, we can conclude the following.

THEOREM 2.6. *An equilibrium point* $u = u(\cdot, u_0)$ *of* (5) *is hyperbolic if and only if* L_y *transversally intersects* $\bar{L}_y$ *for any* $y \in [0, 1]$. *Moreover, if* $W^u(u)$ *denotes the unstable manifold of* $u = u(\cdot, u_0) \in \mathcal{E}$ *and* u *is hyperbolic, then*

$$\dim W^u(u) = 1 + \left[\frac{\Omega(y, u_0)}{\pi}\right]$$

at any $y \in [0, 1]$, *where* $[\cdot]$ *represents the integer part. Consequently,* $u = u(\cdot, u_0) \in \mathcal{E}$ *is hyperbolic and asymptotically stable if and only if* $\Omega(y, u_0) < 0$ *at any* $y \in [0, 1]$.

Proof. Observe that $\Omega(1, u_0) = \theta(u_0) - q_1$. Let us prove first that if the intersection is not transversal at a certain $y \in [0, 1]$, then Ω is constant in y.

We can rewrite the equation for Ω

$$\Omega_y = \frac{\sin^2 \phi - \sin^2 \bar{\phi}}{a(x)} + f_u(x, u)(\cos^2 \phi - \cos^2 \bar{\phi})$$

in the following way:

$$\Omega_y = \left[f_u(x, u) - \frac{1}{a(x)}\right][\cos^2 \phi - \cos^2(\phi - \Omega)].$$

If for any point y, $\Omega(y, u_0) = k\pi$, then $\Omega_y = 0$. Thus the angle Ω will be constant in y. But then $\theta(u_0) = k\pi + q_1$, which implies that u is not hyperbolic, against our assumptions. Consequently,

$$\left[\frac{\Omega(y, u_0)}{\pi}\right] = \left[\frac{\Omega(1, u_0)}{\pi}\right]$$

for any $y \in [0, 1]$. Now we can get the rest of the stated conclusion by applying Theorem 2.2. ∎

Remark. For the nonhomogeneous case, we cannot define an angle similar to σ. But the variational equation (9) remains unchanged and (11) is still the eigenvalue problem in this case. Thus we can still claim the thesis of Theorems 2.2 and 2.6.

On the other hand, for Dirichlet boundary conditions, the maximum principle again tells us that the equilibria are in the interval $[0,1]$. Thus if $\alpha_0 = 1$ or $\alpha_1 = 1$ then we will assume that $\beta_0 \in [0,1]$ or $\beta_1 \in [0,1]$, respectively.

If $\alpha_0 = 1$, we should change the initial values for our shooting method to

$$u(0) = \beta_0, \qquad v(0) = v_0. \tag{14}$$

Then v_0 would substitute u_0 as our "shooting" parameter (which would imply that the initial values for the variational equation (9) now would be $\eta(0) = 0$, $\mu(0) = 1$). The results presented in this section would still hold true for $q_0 = -\pi/2$.

If $\alpha_1 = 1$, we would have to impose an initial value like

$$u(1) = \beta_1, \qquad v(1) = v_1$$

to the backward initial value problem (13). Then $q_1 = \pi/2$ and the results presented in this section would be unchanged otherwise.

In the next section we are going to study a case in which $c(x)$ is a discontinuous function. That will not alter Propositions 2.1 and 2.5 or Theorems 2.2 and 2.6, because they are based on comparisons between angle equations.

3. Discrete space dependence. Let us now assume that $c = c(x) : [0,1] \to \mathbb{R}$ is a step function defined in the following way: c takes the value $c_i \in (0,1) \setminus \{1/2\}$ in the intervals $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, m$; where $0 = x_0 < x_1 < \cdots < x_m < x_{m+1} = 1$.

In each interval $[x_i, x_{i+1}]$, the boundary value problem (4) is a Hamiltonian system, so the orbits of the stationary solutions correspond to the level curves of $H = H(u_0, u_1, c_i)$:

$$H(u_0, u_1, c_i) = \frac{1}{2}(u')^2 + \int_{u_0}^{u_1} u(1-u)(u-c_i)\,du. \tag{15}$$

The linearization of (4) in each interval $[x_i, x_{i+1}]$ is given by the matrix

$$\begin{pmatrix} 0 & 1 \\ -f_u(x,u) & 0 \end{pmatrix},$$

so the eigenvalues $\lambda(u)$ and eigenvectors at $u = 0$ and $u = 1$ are

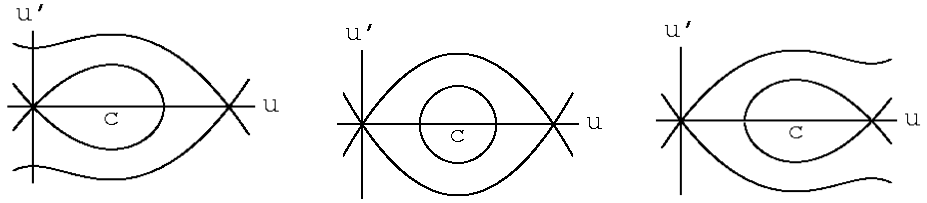
$$\lambda(0) = \pm\sqrt{c_i}: \quad (1, \pm\sqrt{c_i})^T, \tag{16}$$

$$\lambda(1) = \pm\sqrt{1-c_i}: \quad (1, \pm\sqrt{1-c_i})^T. \tag{17}$$

The typical phase portraits (u, u') for different values of c_i are shown in Figure 2. As we can see, for $c_i = 1/2$, there is a heteroclinic orbit connecting the saddle points 0 and 1. For other values of c_i, there is a homoclinic orbit connecting 0 (resp., 1) with itself if $c_i < 1/2$ (resp., $c_i > 1/2$). We will label $\gamma(c_i)$ the point at which the homoclinic orbit crosses the axis $u' = 0$.

It is easy to compute $\gamma(c)$ as a function of c by using the expression (15). In fact, it can be shown that

$$\gamma(c) = \begin{cases} (2 + 2c - \sqrt{4 - 10c + 4c^2})/3 & \text{if } c < 1/2, \\ (-1 + 2c + \sqrt{2}\sqrt{-1 + c + 2c^2})/3 & \text{if } c > 1/2. \end{cases} \tag{18}$$



(a) $c < 1/2$ (b) $c = 1/2$ (c) $c > 1/2$

FIG. 2. *Phase portraits for different values of c.*

We now want to discuss the existence and stability implications of boundary layers at $x = 0$ and $x = 1$. If α_0 and α_1 are different from 1 and ϵ is small enough, the situation is basically the same as that of homogeneous Neumann conditions. We can see this by making the change $x = \epsilon y$, in which case the left boundary condition becomes

$$\epsilon\,\alpha_0\,u(0) - (1-\alpha_0)\,u_y(0) = \epsilon\,\beta_0.$$

Therefore, as in the Neumann case, we can have solutions of (4) which do not exhibit any boundary layer; we also can have solutions with a boundary layer if $u_0 = \gamma(c_1)$, as it is obvious from Figure 2 (namely, the boundary layer would be the upper half of the homoclinic solution).

Nevertheless, the presence of a boundary layer will make the solution unstable.

PROPOSITION 3.1. *If $\alpha_0 \neq 1$ (resp., $\alpha_1 \neq 1$), for any $\delta > 0$ there is $\epsilon_0 > 0$ such that for $\epsilon \in (0, \epsilon_0)$ all the equilibria $u = u(\cdot, u_0) \in \mathcal{E}$ with $u_0 \in [\delta, 1-\delta]$ (resp., $u(1, u_0) \in [\delta, 1-\delta]$) are unstable.*

Proof. Let us assume $\alpha_0 \neq 1$ and $c_1 > 1/2$. (The case $c_1 \leq 1/2$ can be discussed in a similar way.) In order to analyze what happens in the left boundary layer, let us make the change of variables $x = \epsilon y$ in which case we are dealing with the initial value problem:

$$\begin{cases} u_{yy} + u(1-u)(u-c_1) = 0, & 0 < y < x_1/\epsilon, \\ \epsilon\,\alpha_0 u(0) - (1-\alpha_0)\,u_y(0) = \epsilon\,\beta_0. \end{cases} \tag{19}$$

Since $u(0)$ must be a number between 0 and 1, if ϵ is very small, then $u_y(0)$ must be very small in order to verify the initial condition of (19). Also, the initial conditions of the variational equation (9) are going to be

$$\eta(0) = 1, \qquad \mu(0) = \epsilon\,\alpha_0/(1-\alpha_0).$$

Thus, for ϵ small enough, $\mu(0)$ is going to be as small as we wish.

Following Rocha [12], we are going to denote by U a small neighborhood of $(1, 0)$, and by Γ_u and Γ_s the connected components of the unstable and stable manifolds of $(1, 0)$ inside U. The *λ-lemma* shows us (see [10]) that, for $y = x_1/\epsilon$ with ϵ small enough, the curve L_y has a nonempty intersection with U and that $L_y \cap U$ and Γ_u are C^k-close manifolds for any $k \geq 1$. Let us denote by (ζ_0, ζ_1), with $\zeta_0 = \zeta_0(\epsilon)$ and $\zeta_1 = \zeta_1(\epsilon)$, the interval of initial values corresponding to the set $L_y \cap U$ that includes $\gamma(c_1)$.

Now, it is clear from (10) that for $u_0 \in (\zeta_0, \zeta_1)$, we have $\phi(y, u_0) > \pi/2$ and consequently, by Theorem 2.2, the stationary solution $u(x, u_0)$ is unstable. On the order hand, the solutions with $u_0 \in [\delta, \zeta_0]$ leaves the interval $[0, 1]$ for $x \in [0, x_1)$ if ϵ is small enough. Finally, the solutions with $u_0 \in [\zeta_1, 1 - \delta]$, wind around $(c_1, 0)$ and therefore cannot be stable ($\phi(y, u_0) > \pi/2$, again). □

The situation for Dirichlet boundary conditions ($\alpha_0 = 1$ or $\alpha_1 = 1$) is richer, because a solution must have boundary layers if β_0 and β_1 are not 0 or 1. Actually, if $\alpha_0 = 1$, for any $\beta_0 \in (0, 1)$, we may have solutions that exhibit a boundary layer on the left end point of the interval that goes from β_0 to 0 (resp., from β_0 to 1) if $c_1 \geq 1/2$ (resp., $c_1 \leq 1/2$); namely, a piece of the stable manifold of 0 (resp., 1). Analogous things can happen at the right end of the interval $(0, 1)$ but in the reverse sense. In addition to this, if $c > 1/2$ (resp., $c < 1/2$), we can also have a left boundary layer that is a piece of the corresponding homoclinic orbit and starts at β_0 and ends at 1 (resp., 0), or in the reverse sense for a right boundary layer. These layers may exist if we have the following.

1. *Condition* D0: For a left boundary layer,
   (i) $c_1 > 1/2$ and $\gamma(c_1) < \beta_0 < 1$, or
   (ii) $c_1 < 1/2$ and $0 < \beta_0 < \gamma(c_1)$.
2. *Condition* D1: For a right boundary layer,
   (i) $c_m > 1/2$ and $\gamma(c_1) < \beta_1 < 1$, or
   (ii) $c_m < 1/2$ and $0 < \beta_1 < \gamma(c_1)$.

They can be monotone or reach $\gamma(c_1)$ before getting at 1 (resp., 0), but now, only the nonmonotone layers introduce instability.

LEMMA 3.2. *If $u \in \mathcal{E}$ is such that it has a nonmonotone boundary layer at $x = 0$ or $x = 1$ or both, then u is unstable.*

Proof. As it was discussed in the proof of Proposition 3.1, just observe that if u has to wind around the homoclinic orbit, then the angle $\phi(x, u_0)$ corresponding to the variational equation (9) would be greater than $\pi/2$, so the solution $u = u(x, u_0)$ would be unstable. □

Now let us analyze what happens in the interior part of the interval $(0, 1)$. At the beginning, after a possible boundary layer, the solutions that we are considering are located close to $u = 0$ or $u = 1$. Let us study the case of a solution that jumps between 0 and 1, n times in $(0, 1)$. Let us denote by $\{\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_n\} \subset \{x_1, x_2, \ldots, x_m\}$ the jumping points. Let us also call $\bar{x}_0 := 1$ and $\bar{x}_{n+1} := 1$. Then we can prove the following result.

THEOREM 3.3. *There is an $\epsilon_0 > 0$ such that for every $\epsilon \in (0, \epsilon_0)$ the number N_n of stable solutions of* (1) *follows a Fibonacci sequence (that is, $N_k = N_{k-1} + N_{k-2}$, $k = 2, 3, \ldots$) that starts with*

1. *$N_0 = 2$, $N_1 = 3$; if $\alpha_0 \neq 1$ and $\alpha_1 \neq 1$ or if $\alpha_i = 1$ then the condition* Di *is verified, for $i = 0, 1$.*
2. *$N_0 = 1$, $N_1 = 2$; if either*
   (i) *$\alpha_0 = 1$ and the condition* D0 *is not verified but if $\alpha_1 = 1$ then condition* D1 *is verified; or*
   (ii) *$\alpha_1 = 1$ and the condition* D1 *is not verified but if $\alpha_0 = 1$ then condition* D0 *is verified.*
3. *$N_0 = 1$, $N_1 = 1$; if $\alpha_0 = \alpha_1 = 1$ and neither* D0 *nor* D1 *are verified.*

Furthermore, as $\epsilon \downarrow 0$, u approaches 0 or 1 in each open interval $(\bar{x}_i, \bar{x}_{i+1})$ with $i \in \{0, \ldots, n\}$ and it has a monotone transition layer at $\bar{x}_i$ only if

$$[c(\bar{x}_i^+) - c(\bar{x}_i^-)]\, u_x(\bar{x}_i) < 0.$$

Proof. As before, let us assume that $c_1 > 1/2$. The other case can be dealt with by a similar analysis. Also, we are going to consider only the subcase (i) in 2 because we can apply the same analysis to (ii) by going backwards in time.

In situations 2 and 3 of Theorem 3.3, we have a unique stable solution that may exhibit a left boundary layer that goes from β_0 to 0 or may just stay around zero, depending on the initial condition. If there are no changes in the value of $c = c(x)$, there cannot be other stable solutions, so $N_0 = 1$ (in both cases, there may be an end-point boundary layer from 0 to β_1). In case 1, in addition to the previous one, we have another stable solution that may exhibit a left boundary layer that goes from β_0 to 1 monotonically or just remains close to 1, depending also on the initial condition. Thus $N_0 = 2$ in this case.

We can now consider a small neighborhood V of $(0,0)$ similar to U, the small neighborhood of $(1,0)$ used in the proof of Proposition 3.1. If at $x_1 < 1$ the function c jumps to a value $c_2 > 1/2$, no new stable solutions are introduced. But if $c_2 < 1/2$, then we can have a stable solution that exhibits an internal layer from 0 to 1, because the new stable manifold of the equilibrium $(1,0)$ has a transversal intersection to the previous unstable manifold of the origin (and the point of intersection is unique). The λ-lemma shows us that for $y = \bar{x}_2/\epsilon$, $L_y \cap U$ is nonempty and it is C^k-closed to $\Gamma^u \cap U$, the piece of the unstable manifold of $(1,0)$ inside U.

This solution with an internal transition will be stable because the angle corresponding to its variational equation is negative. (Observe that $L_y \cap U$ intersect transversally with $\bar{L}_y \cap U$.)

If there are no more jumps in c, we have that $N_2 = 3$ if we are in situation 1, and $N_2 = 2$ if we are in situation 2; in both cases, there actually exists a solution that exhibits an internal layer because it can satisfy any boundary condition at the end point either by staying at 1 or by having a left homoclinic boundary layer to β_1 (by the hypothesis in 2, if we have a Dirichlet end-point condition, the line $x = \beta_1$ must cross the homoclinic orbit of $(1,0)$). But there is no way of getting to β_1 from 0 in case 3. Thus the only stable solution in 3 is the one with the internal transition from 0 to 1.

We can proceed doing the same analysis at every jumping point. Actually, at each $\bar{x}_k$, there will be a new transversal intersection between the unstable manifold of one of the equilibrium points $(0,0)$ or $(1,0)$, and the stable manifold of the other one. This allows the solutions that are close to the first equilibrium point either to stay there or cross to the other equilibrium point without losing their stability.

Thus let us denote by N_k^U and N_k^V the number of intersections existing at the jump point $\bar{x}_k$ in the neighborhoods U and V, respectively. Let us assume that the jump crosses $1/2$ positively, that is, $c(\bar{x}_k^+) - c(\bar{x}_k^-) > 0$. Then

$$\begin{aligned} N_{k+1}^V &= N_k^V + N_k^U, \\ N_{k+1}^U &= N_k^U. \end{aligned}$$

Moreover, since the jump at $\bar{x}_{k-1}$ has the opposite sign, we have

$$\begin{aligned} N_k^V &= N_{k-1}^V, \\ N_k^U &= N_{k-1}^U + N_{k-1}^V. \end{aligned}$$

Now, in cases 1 and 2, we can verify the end-point boundary condition either from 0 or 1. Thus $N_k = N_k^U + N_k^V$, and then we have

$$N_{k+1} = N_{k+1}^U + N_{k+1}^V = N_k^U + N_k^V + N_k^U = N_k + N_{k-1}^U + N_{k-1}^V.$$

Consequently, $N_{k+1} = N_k + N_{k-1}$. In case 3 only the solutions that are close to 0 after the jump $k+1$ will be able to verify the end-point boundary condition. Thus

$$N_{k+1} = N_{k+1}^V = N_k^V + N_k^U = N_{k-1}^V + N_k = N_{k-1} + N_k,$$

and this finishes the proof. □

We can also study the existence of solutions that exhibit nonmonotone boundary layers. They are going to exist in cases 1 and 2 of Theorem 3.3. Their *Morse index* (dimension of the unstable manifold) is at least 1, due to the nonmonotonicity of the boundary layer. In case 1 of Theorem 3.3, we can even have solutions that exhibit two nonmonotone boundary layers (consequently, their Morse index will be at least 2).

4. Some examples. It is very difficult to determine the existence and stability properties of the unstable solutions of (1) in the same way that we did with the stable ones in Theorem 3.3, because in these problems there are several sources of unstable solutions. Thus we are going to study some particular examples taken from [11], in order to determine completely both the attractors and the flow in them.

The first example analyzed in [11] has as $c(x)$ the step function ($n = 1$)

$$c(x) = \begin{cases} 1/2 + c_1, & x < c_3, \\ 1/2 - c_2, & x \geq c_3, \end{cases} \tag{20}$$

where $c_1, c_2 \in (0, 1/2)$ and $c_3 \in (0, 1)$. Thus $c(x)$ has a negative jump at $x = c_3$ from $1/2 + c_1$ to $1/2 - c_2$. A similar example can be constructed by using a $\bar{c}(x) = c(-x)$ which exhibit a positive jump.

Across this jump at $x = c_3$, the phase diagram changes from having a homoclinic orbit around $u = 1$ to having another one around $u = 0$. As we did before, we are going to denote the intersection of the homoclinic orbit, that exists around zero for $x \geq c_3$, with the u-axis by $\gamma(c_2)$. But for $x < c_3$, we will represent the corresponding intersection by $1 - \gamma(c_1)$. Thus if we have that $\gamma(c_1) + \gamma(c_2) \leq 1$ then there is at most a contact point between the two homoclinic orbit. Having this in mind, we state the following theorem.

THEOREM 4.1. *Consider the equation* (1)*, with boundary conditions* (2) *and* $c(x)$ *defined as in* (20)*. If* c_1 *and* c_2 *satisfy the condition* $\gamma(c_1) + \gamma(c_2) \leq 1$*, then there exists* ϵ_0 *such that for every* $\epsilon \in (0, \epsilon_0)$ *the attractor of this problem can be represented by a linear segment* (*see Figure* 3) *and is described as follows:*

1. *For the case in which* $\alpha_0 \neq 1$ *and* $\alpha_1 \neq 1$ *or if* $\alpha_i = 1$*, then the condition* Di *is verified for* $i = 0, 1$*; we have that the attractor has exactly* 5 *stationary solutions:* 3 *stable ones and* 2 *unstable ones. The unstable ones have Morse index* 1*, and the unstable manifolds of each unstable solution are made of two heteroclinic orbits, one connecting to the equilibrium that exhibits a monotone internal transition from* 0 *to* 1*, and the other to one of the other stable solutions.*

2. *For the case in which either*

(i) $\alpha_0 = 1$ *and the condition* D0 *is not verified, but if* $\alpha_1 = 1$*, then condition* D1 *is verified, or*

(ii) $\alpha_1 = 1$ *and the condition* D1 *is not verified, but if* $\alpha_0 = 1$*, then condition* D0 *is verified;*

the attractor has exactly 3 *stationary solutions:* 2 *stable ones and* 1 *unstable one. One of the stable solutions exhibits a monotone internal transition from* 0 *to* 1*, and the unstable solution has Morse index* 1*. The unstable manifold of the unstable solution*

is made of two heteroclinic orbits, one connecting to the equilibrium that exhibits a monotone internal transition layer from 0 *to* 1*, and the other one to the other stable solution.*

3. *For the case in which* $\alpha_0 = \alpha_1 = 1$ *and neither* D0 *nor* D1 *are verified, the attractor is made of a unique stable solution that exhibits a monotone internal transition from* 0 *to* 1.

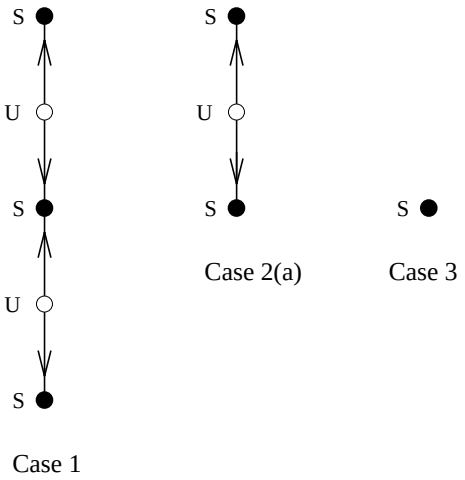


FIG. 3. *Attractors of Theorem* 4.1.

Proof. Let us analyze the last case first. In case 3, we have Dirichlet boundary conditions at both ends of the interval and both of the end points are outside the homoclinic orbits of their corresponding phase portraits (see Figure 4). By the λ-lemma, the time maps L_y and $\bar{L}_y$ are C^k-close to the unstable manifold of 0 before the jump in $c(x)$ and the stable manifold of 1 after the jump, respectively. Thus L_y and $\bar{L}_y$ have only one intersection that is going to be close to the intersections of these invariant manifolds. Consequently, the intersection angle $\Omega(y, v_0)$ is going to be very close to the angle of the intersection of the invariant manifolds.

In order to determine the sign of the intersection angle between the unstable manifold of 0 before the jump in $c(x)$ and the stable manifold of 1 after the jump, let us write down the orbital equation that can be obtained from (7),

$$\frac{dv}{du} = -\frac{f(x,u)}{v}.$$

Thus the difference in the derivatives at an intersection point (u^*, v^*) of the invariant manifolds is

$$\frac{dv_1}{du_1} - \frac{dv_2}{du_2} = -\frac{f_2(x,u^*) - f_1(x,u^*)}{v^*} = \frac{u^*(1-u^*)}{v^*}(c_2 + c_1) > 0. \tag{21}$$

The angle at the intersection point of the two invariant manifolds is negative in the sense defined by Theorem 2.6. Then the stationary solution corresponding to the unique intersection of L_y and $\bar{L}_y$ is stable (thus this is the only stationary solution in this case, which we already knew from Theorem 3.3). This solution exhibits an internal transition layer between 0 and 1.

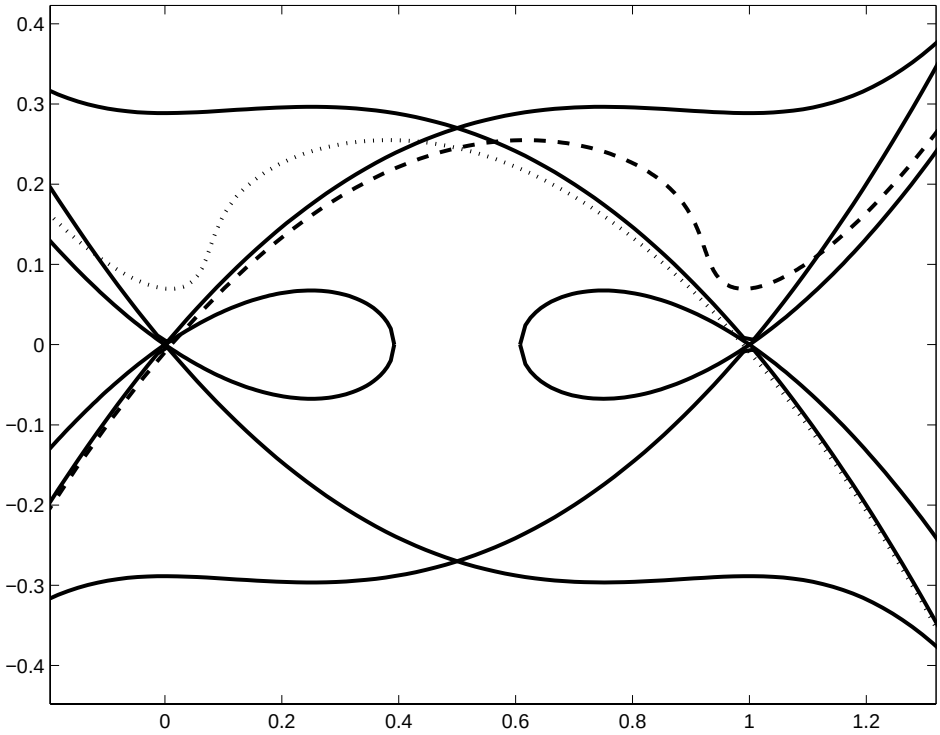


FIG. 4. *Numerical approximations of the time maps L_y (dashed line) and $\bar{L}_y$ (dotted line) in case 3 with $\beta_0 = 0.55$, $\beta_1 = 0.45$, and $c_1 = c_2 = 0.25$.*

This solution also is going to exist in the other cases and is always going to be stable (this can be proven in the same fashion as before). In case 2, say (i), the initial condition is also of Dirichlet type, but the initial point β_0 is now inside the homoclinic orbit that exists around 1 for $c(x) = c_1$. This produces a more complicated geometry for the forward time map. By again using the λ-lemma, we can conclude L_y is still going to be C^k-close to the unstable manifold of 0 before the jump in $c(x)$, but then it will have to curve around, get close to the unstable manifold of 1 before the jump, turn around again inside the homoclinic orbit, and get close again to the unstable solution of 1 before the jump (see Figure 5). This introduces two new crossings between L_y and $\bar{L}_y$ because the backward time map $\bar{L}_y$ is C^k-close to the unstable manifold of 1 for $c(x) = c_2$ (it will include the point $(1, 0)$ if the end-point boundary condition is of Neumann type).

The closest intersection to zero corresponds to a stable equilibrium while the other intersection nearby corresponds to an equilibrium that is unstable (it exhibits a nonmonotone boundary layer). We can see this by using formulas (16) and (17) that directly give us

$$\arctan\sqrt{1/2 - c_1} - \arctan\sqrt{1/2 + c_2},$$

as the angle between the two corresponding invariant manifolds at their intersection point, which is negative in the sense defined by Theorem 2.6. Thus the first intersection corresponds to a stable solution. The second one is unstable because the

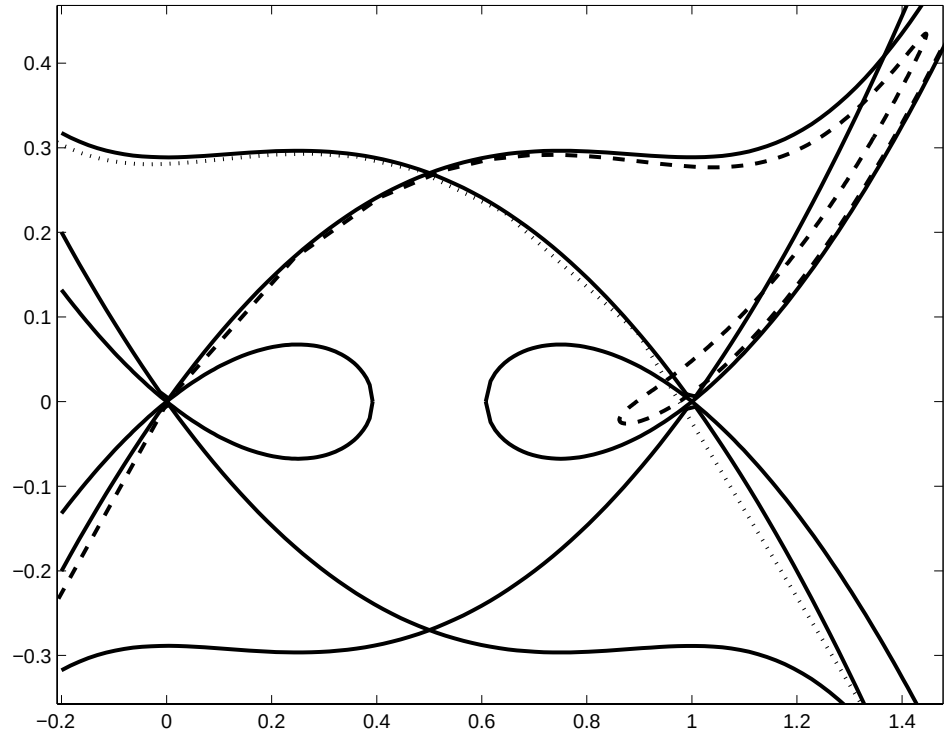


FIG. 5. *Numerical approximations of the time maps L_y (dashed line) and $\bar{L}_y$ (dotted line) in case* 2(i) *with* $\beta_0 = 0.65$, $\beta_1 = 0.45$, *and* $c_1 = c_2 = 0.25$.

intersection angle is increased at least by 2π with respect to the previous one. We could have gotten the same conclusion by analyzing (21).

Finally, in case 1, we have a similar situation to the right side of case 2(i), but now in both ends. Consequently, we have a total of 5 equilibria: 3 stable and 2 unstable (see Figure 6). If the boundary conditions are of Neumann type, then two of the stable intersections between L_y and $\bar{L}_y$ happen at the points $(0,0)$ and $(1,0)$, which are the end points of L_y and $\bar{L}_y$. Otherwise, they happen nearby and they are true crossings.

To determine the connections, we can use the method describe in [4]. In this case, we have to use a permutation matrix formed by the derivatives of the solutions at $x = 0$ and $x = 1$. The computation is straightforward.

As ϵ decreased, the geometry of the time maps inside the homoclinic orbits becomes more complicated. But we know that there are at most two intersections between a time map and the corresponding homoclinic orbit. Outside the homoclinic orbits, the time maps behave much more nicely, approaching the corresponding invariant manifolds of 0 and 1, more and more as $\epsilon \downarrow 0$. □

An example with two jumps in $c(x)$ can be also analyzed. Let $c(x)$ be the step function ($n = 2$)

$$c(x) = \begin{cases} 1/2 + c_1, & x < c_3, \\ 1/2 - c_2, & c_3 \le x < c_5, \\ 1/2 + c_4, & c_5 \le x, \end{cases} \tag{22}$$

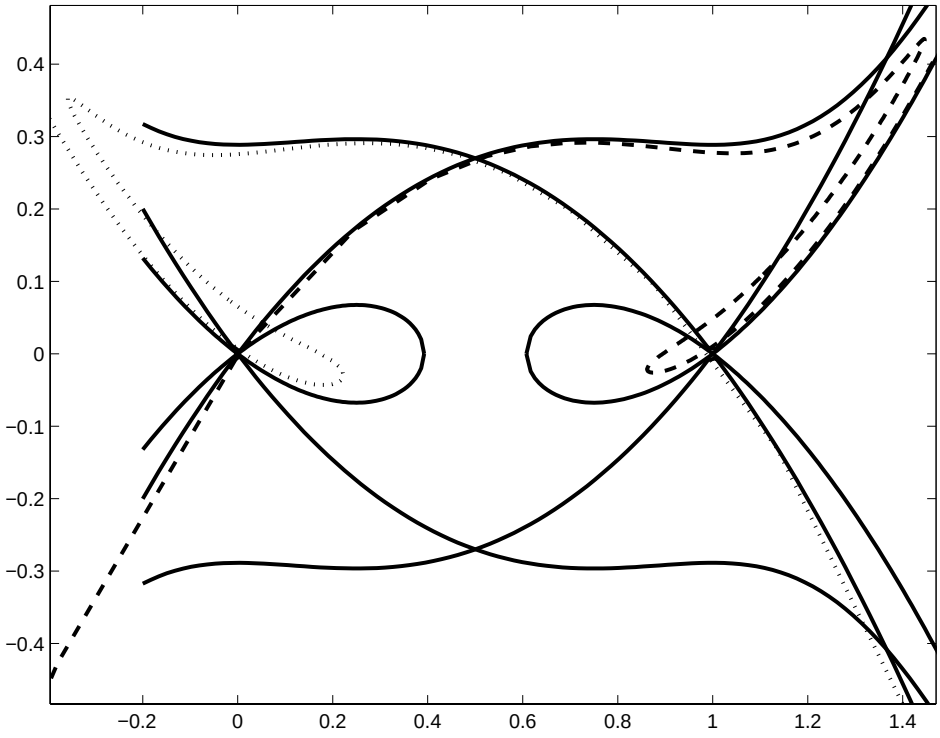


FIG. 6. *Numerical approximations of the time maps L_y (dashed line) and $\bar{L}_y$ (dotted line) in case* 1 *with* $\beta_0 = 0.65$, $\beta_1 = 0.3$, *and* $c_1 = c_2 = 0.25$.

where $c_1, c_2, c_4 \in (0, 1/2)$ and $c_3, c_5 \in (0, 1)$. Therefore $c(x)$ has a negative jump at $x = c_3$ from $1/2 + c_1$ to $1/2 - c_2$ and a positive jump at $x = c_5$ from $1/2 - c_2$ to $1/2 + c_4$. We can now state the following theorem. (Here, $\gamma(c_3)$ is defined as $\gamma(c_1)$ in Theorem 4.1.)

THEOREM 4.2. *Consider* (1), *with boundary conditions* (2) *and* $c(x)$ *defined as in* (22). *If* c_1, c_2, *and* c_4 *satisfy the conditions* $\gamma(c_1) + \gamma(c_2) \leq 1$ *and* $\gamma(c_2) + \gamma(c_4) \leq 1$, *then there exists* ϵ_0 *such that for every* $\epsilon \in (0, \epsilon_0)$ *the attractor of this problem can be represented as in Figure* 7 *and is described as follows:*

1. *For the case in which* $\alpha_0 \neq 1$ *and* $\alpha_1 \neq 1$ *or if* $\alpha_i = 1$ *then the condition* Di *is verified for* $i = 0, 1$; *we have that the attractor has exactly* 11 *stationary solutions:* 5 *stable ones,* 5 *unstable ones of Morse index* 1, *and* 1 *unstable one of Morse index* 2. *The unstable manifolds of the unstable stationary solutions of Morse index* 1 *are made of two heteroclinic orbits which connect them with the stable solutions next to them. The unstable stationary solution of Morse index* 2 *is connected with all the other equilibria except two of them, forming a "racquet"-like attractor* (*see Figure* 7).

2. *For the case in which either*

(i) $\alpha_0 = 1$ *and the condition* D0 *is not verified but if* $\alpha_1 = 1$ *then condition* D1 *is verified, or*

(ii) $\alpha_1 = 1$ *and the condition* D1 *is not verified but if* $\alpha_0 = 1$ *then condition* D0 *is verified,*

we have that the attractor has exactly 5 *stationary solutions:* 3 *stable ones and* 2

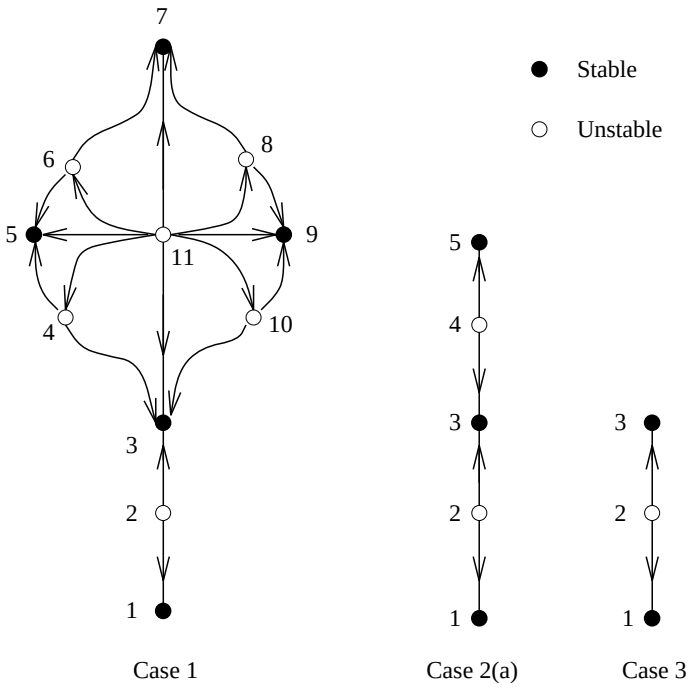


FIG. 7. *Attractors of Theorem* 4.2.

unstable ones. The unstable ones have Morse index 1 *and the unstable manifolds of each unstable solutions are made of two heteroclinic orbits, one connecting to the equilibrium that exhibits two monotone internal transitions (one from* 0 *to* 1 *and the other back from* 1 *to* 0*), and the other one connecting to one of the other stable solutions.*

3. *For the case in which* $\alpha_0 = \alpha_1 = 1$ *and neither* D0 *nor* D1 *are verified; we have that the attractor is made of* 3 *stationary solutions:* 2 *stable ones and* 1 *unstable one. One of the stable ones exhibits two monotone internal transitions (one from* 0 *to* 1 *and the other back from* 1 *to* 0*), and the unstable one has Morse index* 1*. The unstable manifold of the unstable solution is made of two heteroclinic orbits that connect it to the stable solutions.*

Proof. Observe that since now we have two jump points (which, once we make the change of variable $y = x/\epsilon$, will be represented by y_1 and y_2), it may be interesting to look at the intersections between the time maps L_y and $\bar{L}_y$ at the points $y = y_1$ and $y = y_2$. As in the proof of Theorem 4.1, we start now by the simplest case, which is the last one.

Case 3 is very similar to case 2(i) in Theorem 4.1. Because the line $u = \beta_0$ does not intersect the homoclinic orbit around 1 that exists for $x < c_3$, L_{y_1} is C^k-close to the unstable manifold of 0 for ϵ small enough by the λ-lemma. Therefore L_{y_1} does intersect the homoclinic solution that exists around 0 for $c_3 \leq x < c_5$. That means that L_{y_2} will exhibit a "finger" that would go inside the homoclinic orbit around zero. That finger will cross the homoclinic orbit exactly twice and it will have to turn around and join the rest of the curve L_{y_2} that will be now C^k-close to the unstable manifold of 1. On its left side, the finger will leave the interval $[0, 1]$ by being C^k-close to the unstable manifold of 0. See Figure 8.

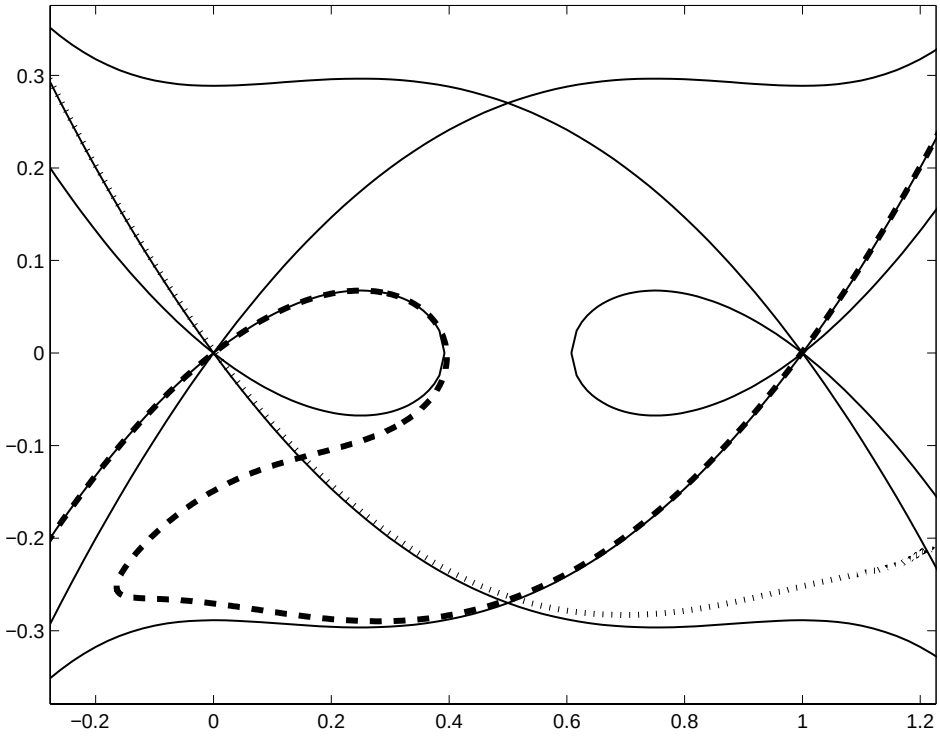


FIG. 8. *Numerical approximations of the time maps L_{y_2} (dashed line) and $\bar{L}_{y_2}$ (dotted line) in case 3 with $\beta_0 = 0.55$, $\beta_1 = 0.45$, and $c_1 = c_2 = c_4 = 0.25$.*

The curve $\bar{L}_{y_2}$, on the other hand, will be C^k-close to the stable manifold of 0 (inside the interval $[0, 1]$). Thus it will intersect L_{y_2} three times. Numbering those intersections in increasing order as one progress along $\bar{L}_{y_2}$ from $u = 0$ to $u = 1$, we have that the first intersection corresponds to a stable solution (the intersection angle gets close to the value

$$\arctan\sqrt{1/2 + c_4} - \arctan\sqrt{1/2 - c_2},$$

as ϵ goes to 0), the second to an unstable solution, and the third to a stable solution. (The angle at this last intersection gets close to the angle of intersection between the stable manifold of 0 for $c_3 \le x < c_5$ and the unstable manifold of 1 for $x > c_5$, which is negative—to see this requires a computation similar to (21).)

The first solution stays close to 0 most of the time but exhibits two monotone boundary layers connecting 0 with β_0 and β_1. The other two stationary solutions exhibit the same boundary layers, but the unstable one exhibits a "bump" between $y = y_1$ and $y = y_2$ (it goes around the homoclinic orbit of 0 for $c_3 \le x < c_5$; that is what makes it unstable) and the other stable solution exhibits two internal jumps: one form 0 to 1 at $y = y_1$ and the other back from 1 to 0 at $y = y_2$. These three solutions also exist in all the other cases.

In case 2 (i), L_{y_1} will also be C^k-close to the unstable manifold of 0 for $x < c_3$, but then it will turn around and have a finger that goes inside the homoclinic orbit around 1 because the line $u = \beta_0$ does now intersect that homoclinic orbit. The time

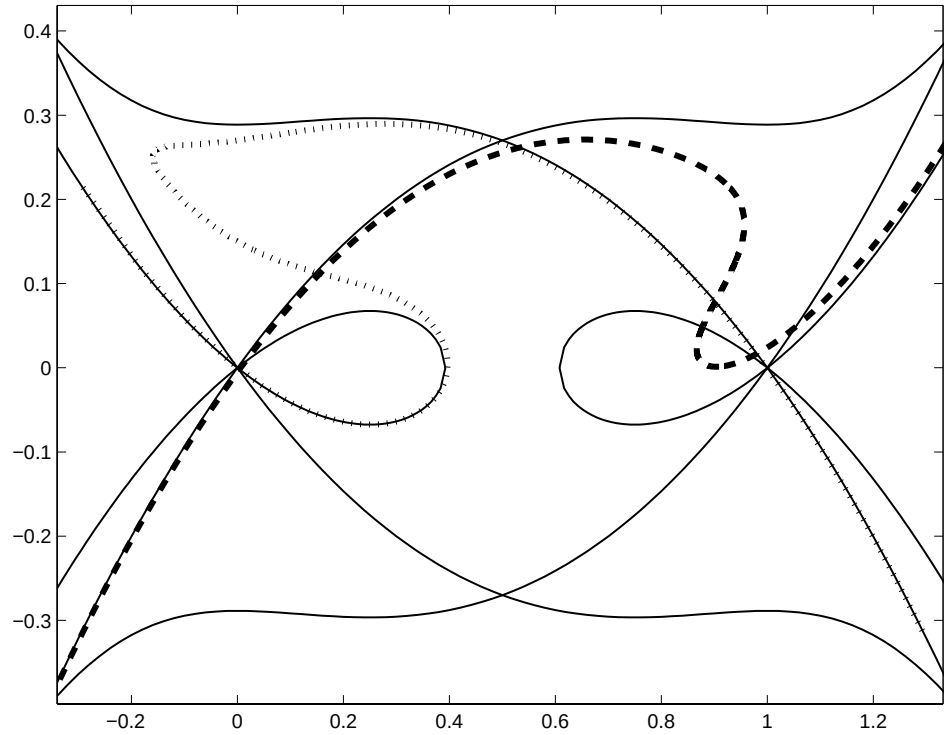


FIG. 9. *Numerical approximations of the time maps L_{y_1} (dashed line) and $\bar{L}_{y_1}$ (dotted line) in case* 2(i) *with* $\beta_0 = 0.65$, $\beta_1 = 0.45$, *and* $c_1 = c_2 = c_4 = 0.25$.

map $\bar{L}_{y_1}$, on the other hand, will have a similar finger that goes inside the homoclinic orbit around 0 that exist for $c_3 \leq x < c_5$ (as in the previous case but backwards in time). This introduces two new crossings. Numbering all intersections consecutively as we progress along $\bar{L}_{y_1}$ from $u = 0$ to $u = 1$, the three first correspond to the same equilibria as that in the previous case. The number 4 is unstable of index 1 (it exhibits a nonmonotone boundary layer at $x = 0$) and the number 5 is stable (similar argument that for number 3 in the previous case). This last equilibrium point exhibits an internal transition layer from 1 to zero at $y = y_2$. See Figure 9.

If we now look at the situation for $y = y_2$, we will see that L_{y_2} has a finger going inside the homoclinic orbit around 0 that exists for $c_3 \leq x < c_5$, then turns around and gets C^k-close to the unstable manifold of 1. But now, it has three pleats along that unstable manifold. They are elongations of the finger that L_{y_1} has close to 1. The five intersections that L_{y_2} and $\bar{L}_{y_2}$ have, when look at along $\bar{L}_{y_2}$ from $u = 0$ to $u = 1$, correspond to the stationary solutions 1, 2, 5, 4, and 3. Thus the order is not preserved. But this intersection (and this ordering) is preserved in case 1, as we can see in Figure 10.

Actually the diagram for case 1 at $y = y_2$ looks very much the same than in the previous case except that now $\bar{L}_{y_2}$ also has a finger that goes inside the homoclinic orbit that exists around 1 for $x \geq c_5$ (see Figure 10).

This new finger introduces six new intersections. The properties and layers of these new equilibria can be deduced by looking at which side of the finger in $\bar{L}_{y_2}$ and

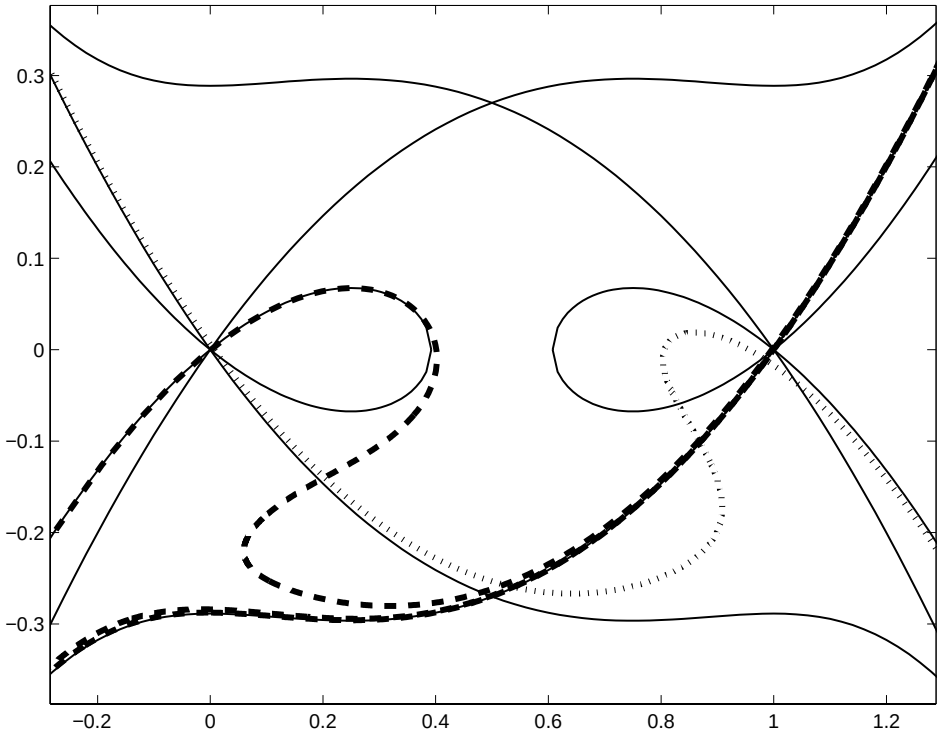


FIG. 10. *Numerical approximations of the time maps L_{y_2} (dashed line) and $\bar{L}_{y_2}$ (dotted line) in case* 1 *with $\beta_0 = 0.65$, $\beta_1 = 0.7$, and $c_1 = c_2 = c_4 = 0.25$. L_{y_2} folds three times along the unstable manifold of* 1.

which pleat of L_{y_2} they are. For example, along the same pleat as solution number 5 we find two new intersections, corresponding to solution numbers 6 and 7. That means that until $y = y_2$, solutions 5, 6, and 7 behave very much alike. But while solution number 5, being outside the finger, makes a transition at $y = y_2$ from 1 to zero, solution number 6 exhibits a nonmonotone layer from 1 to β_1 (becoming unstable of Morse index 1 as a consequence), and solution number 7 exhibits a monotone layer from 1 to β_1 (resulting in a stable solution; this can be checked in a similar fashion as before).

In the same side of the finger as solution number 7 we find solution numbers 8 and 9. Because they occupy such positions, solution numbers 8 and 9 also exhibit monotone boundary layers from 1 to β_1. But they are in the two consecutive pleats of L_{y_2} after solution number 7, so solution number 8 exhibits a nonmonotone boundary layer from β_0 to 1 (becoming unstable of index 1) and solution number 9 exhibits a monotone layer from β_0 to 0; it follows an internal transition layer from 0 to 1 at $y = y_1$. Thus solution number 9 is stable.

On the same pleat as solution number 9 on the other side of the finger we find solution number 10. The only difference between these two solutions is that the left boundary layer of solution number 10 is nonmonotone. Consequently, solution number 10 is unstable of index 1.

Finally, on the second pleat of L_{y_2} and on the left side of the finger of $\bar{L}_{y_2}$, we find solution number 11. From this location we conclude that solution number 11 must

behave like solution numbers 4 and 8 before $y = y_2$ (therefore having a nonmonotone left boundary layer) and like solution numbers 6 and 10 afterwards (therefore having a nonmonotone right boundary layer). Thus solution number 11 must be unstable of Morse index 2.

The method given in [4] allows us again to prove that solution number 11 is connected to all the other solutions except numbers 1 and 2. The cascading property then implies all the other stated connections. □

5. Persistence of solutions after smoothing *c*. We would like to prove that close to every stationary solution $u = u(x)$ of the discrete case—where $c = c(x)$ is a step function—there is a stationary solution $\tilde{u} = u + v$ for the continuous case $\tilde{c} = c + \chi^\delta$. We also would like to prove that the stability properties of u are preserved.

Consequently we want v to verify

$$\begin{cases} \epsilon^2 D^2 v + g^\epsilon(\delta, v) = 0 & \text{in } (0,1), \\ \alpha_0\, v(0) - (1-\alpha_0)\, v'(0) = 0, \\ \alpha_1\, v(1) + (1-\alpha_1)\, v'(1) = 0, \end{cases} \tag{23}$$

where D^2 stands for the second derivative and

$$g^\epsilon(\delta, v) = \epsilon^2 D^2 u + f^\delta(x, u+v),$$

with

$$f^\delta(x,u) = u(1-u)(u - c(x) - \chi^\delta).$$

In order to use the Implicit Function theorem (IFT), let $\epsilon \in (0, \epsilon_0)$, and let us define the operator $F^\epsilon : [0,\delta_0] \times W_0^{2,\,p}([0,1]) \longrightarrow L^p([0,1])$ as

$$F^\epsilon(\delta, v) = \epsilon^2 v + D^{-2} g^\epsilon(\delta, v).$$

Observe that if v belongs to $W_0^{2,\,p}([0,1])$, it verifies the boundary conditions in (23). Also notice that we can always invert the operator D^2 in $W_0^{2,\,p}([0,1])$.

For simplicity's sake, let us study the case in which c has only one jump at the point $x = x_1$. Then c will look like

$$c(x) = \begin{cases} c_0 > 1/2, & x \in [0, x_1], \\ c_1 < 1/2, & x \in (x_1, 1]. \end{cases}$$

The functional χ^δ is a perturbation of c such that

$$\chi^\delta(x) = \chi\left(\frac{x - x_1}{\delta}\right),$$

where $\chi \in Z = L^p(-\infty, +\infty)$. Actually, we are particularly interested in a perturbation like

$$\chi(x) = \begin{cases} -Je^x, & x \in (-\infty, 0), \\ Je^{-x}, & x \in [0, +\infty) \end{cases} \tag{24}$$

for $J = (c_0 - c_1)/2$ (see Figure 11) that would make $c + \chi^\delta$ a continuous function, although we will treat all perturbations in $Y = L^p([0,1])$.

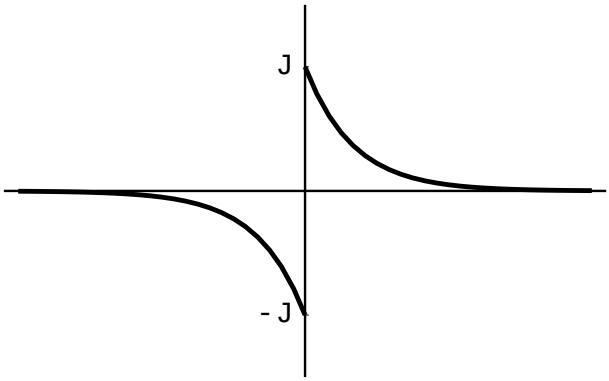


FIG. 11. *Desired perturbation function* $\chi(x)$.

The functional $g^\epsilon(\delta, v) : [0, \delta_0] \times L^p([0,1]) \longrightarrow L^p([0,1])$ is C^∞ with respect to v and is Hölder continuous of exponent $1/p$ with respect to δ, because

$$\|\chi^\delta\|_Y^p = \int_0^1 \left|\chi\left(\frac{x - x_1}{\delta}\right)\right|^p \mathrm{d}x = \int_{-\frac{x_1}{\delta}}^{\frac{1-x_1}{\delta}} |\chi(y)|^p \, \delta \mathrm{d}y \leq \delta \|\chi\|_Z^p.$$

Since F^ϵ is then Hölder continuous of exponent $1/p$ with respect to δ and differentiable with respect to v, $F^\epsilon(0,0) = 0$, and F_v^ϵ is invertible because u is hyperbolic, we can apply the IFT and obtain $\delta_1 = \delta_1(\epsilon) > 0$, such that for every $\delta \in [0, \delta_1)$, $\tilde{u}(\delta) = u + v(\delta)$ is a stationary solution of (1) that is close to u and is a Hölder continuous function of δ of exponent $1/p$.

Now, since $v(\delta)$ belongs to the space $W_0^{2,\,p}$, then by the general Sobolev inequalities, $v \in C^{1,\,q}([0,1])$ with $q = 1 - 1/p$ if $p > 1$ and $v(\delta) \in C^{0,\,q}([0,1])$ for any $0 < q < 1$ if $p = 1$. Thus if in addition to $p > 1$, we choose the perturbation χ^δ such that $c + \chi^\delta$ is a continuous function—selecting it as in (24), for example—then $\tilde{u} \in C^2([0,1])$.

We can execute a similar procedure for every discontinuous term c. We will represent by $\mathcal{E}(0) = \mathcal{E}$ the set of stationary solution for each c and by $\mathcal{E}(\delta)$ the set of stationary solutions that arise from the perturbation χ^δ.

Due to the continuity of the functional χ^δ with respect to δ, the eigenvalues of the linearization of (1) around $\tilde{u}$ are continuous functions of δ. Thus there exists a $\delta_2 = \delta_2(\epsilon) > 0$ such that for $\delta \in [0, \delta_2)$, the stability properties of $\tilde{u}$ are the same as u.

Finally, we want to prove that there exists a $\delta_3 > 0$, such that for $\delta \in [0, \delta_3)$, all the stationary solutions of (1) are in $\mathcal{E}(\delta)$. For that, we will need the following result.

LEMMA 5.1. *Let u be a stationary solution of* (1) *for $\epsilon \in (0, \epsilon_0)$. Then $|u(x)| \leq 1$ for all $x \in [0,1]$ and if $\exists x \in [0,1]$ such that $u(x) = 0$ (resp., $u(x) = 1$), then u is constant equal to* 0 *(resp., equal to* 1*) in* $[0,1]$.

Proof. Let us assume that $\exists x_0 \in [0,1]$ such that $u(x_0) = 0$. If $u'(x_0) = 0$, then by uniqueness of solutions of the initial value problem, $u \equiv 0$.

If $u'(x_0) < 0$, then in order for u to be a stationary solution, $u''(x) < 0$ for all $x \in (x_0, 1]$ because $f(x,u) > 0$ for all $u < 0$. Then u could not arrive at $x = 1$ to a value in $[0,1]$. This implies that for ϵ_0 small enough, u cannot be a stationary solution. The other case could be discussed similarly. □

Let us now assume that there is no such δ_3. Then we can construct a sequence $\{\delta_n\}_{n=1}^\infty$ such that for every δ_n, there is a stationary solution $u_n \notin \mathcal{E}(\delta_n)$. By the IFT,

this implies that all the u_n are outside the uniqueness neighborhoods of the functions in $\mathcal{E}(\delta_n)$. For simplicity's sake, let us assume that u_n is a classical solution, that is, $u_n \in C^2$, although the following argument would still be true with weaker smoothness assumptions after making the appropriate corrections.

Consequently, $\{u_n\}_{n=1}^{\infty}$ is a sequence of continuous functions defined in $[0,1]$, which is a compact subset of $\mathbb{R}$. This sequence is equibounded by Lemma 5.1. Their first and second derivatives are also equibounded, because

$$|u_n''| \leq \frac{|f(x,u_n)|}{\epsilon^2} \leq M(\epsilon).$$

Thus both $\{u_n\}_{n=1}^{\infty}$ and $\{u_n'\}_{n=1}^{\infty}$ converge uniformly when $n \to \infty$, by the Arzelà–Ascoli theorem. The limit function u_0 is a C^2 stationary solution that is not in $\mathcal{E}$. This is a contradiction. Therefore, taking $\delta_0(\epsilon) = \min\{\delta_1, \delta_2, \delta_3\}$, we have proved the following.

THEOREM 5.2. *There exists an $\epsilon_0 > 0$ such that for every $\epsilon \in (0, \epsilon_0)$, there exists a $\delta_0 = \delta_0(\epsilon)$ such that for every $\delta \in [0, \delta_0)$, the attractor $\mathcal{A}$ of (1) for a discrete c and the attractor $\mathcal{A}_\delta$ of (1) for $c + \chi_\delta$ are topologically equivalent. That is, $\mathcal{A} \cong \mathcal{A}_\delta$.*

This theorem also could have been deduced from more general results proven in [5].

REFERENCES

[1] S. B. ANGENENT, J. MALLET-PARET, AND L. A. PELETIER, *Stable transition layers in a semilinear boundary value problem*, J. Differential Equations, 67 (1987), pp. 212–242.

[2] G. BIRKHOFF AND G.-C. ROTA, *Ordinary Differential Equations*, 2nd ed., Blaisdell Publishing Company, New York, 1969.

[3] B. FIEDLER AND C. ROCHA, *Orbit equivalence of global attractors of semilinear parabolic differential equations*, Trans. Amer. Math. Soc., to appear.

[4] B. FIEDLER AND C. ROCHA, *Heteroclinic orbits of semilinear parabolic equations*, J. Differential Equations, 125 (1996), pp. 239–281.

[5] J. K. HALE AND G. RAUGEL, *Lower semicontinuity of attractors of gradient systems and applications*, Ann. Mat. Pura Appl. IV, CLIV (1989), pp. 281–326.

[6] J. K. HALE AND K. SAKAMOTO, *Existence and stability of transition layers*, Japan J. Appl. Mathematics, 5 (1988), pp. 367–405.

[7] H. L. KURLAND, *Monotone and oscillatory equilibrium solutions of a problem arising in population genetics*, in Nonlinear Partial Differential Equations, Contemp. Math., 17, American Mathematical Society, Providence, RI, 1983, pp. 323–342.

[8] J. KWAPISZ, *Stability in a semilinear boundary value problem via invariant conefields*, J. Differential Equations, 141 (1997), pp. 86–101.

[9] H. MATANO, *Convergence of solutions of one-dimensional semilinear parabolic equations*, J. Math. Kyoto Univ., 18–2 (1978), pp. 221–227.

[10] J. PALIS, JR. AND W. DE MELO, *Geometric Theory of Dynamical Systems*, A. K. Manning, trans., Springer-Verlag, New York, Berlin, 1982.

[11] C. ROCHA, *Generic properties of equilibria of reaction-diffusion equations with variable diffusion*, Proc. Roy. Soc. Edinburgh Sect. A, 101 (1985), pp. 45–55.

[12] C. ROCHA, *Examples of attractors in scalar reaction-diffusion equations*, J. Differential Equations, 73 (1988), pp. 178–195.

[13] E. YANAGIDA, *Stability of stationary distributions in a space-dependent population growth process*, J. Math. Biol., 15 (1982), pp. 37–50.

[14] T. I. ZELENYAK, *Stabilization of solutions of boundary value problems for a second order parabolic equation with one space variable*, Differential Equations, 4 (1968), pp. 17–22.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 985–1009

A FREE BOUNDARY PROBLEM FOR SCALAR CONSERVATION LAWS*

ANDREA TERRACINA†

Abstract. We investigate a class of free boundary problems for scalar conservation laws, which is suggested by a model of ion etching. First we give an entropy formulation of the problem; then we prove existence, uniqueness, continuous dependence, and comparison properties of solutions for a large class of initial data.

Key words. conservation laws, free boundary

AMS subject classifications. 35L65, 35R35

PII. S0036141097325307

1. Introduction. There is a very rich literature about Stefan problems for parabolic and elliptic equations and systems (e.g., see [Da] and references therein). On the other hand, free boundary problems for hyperbolic equations and systems are much less investigated.

Nevertheless, many physical models (for instance, in gas dynamics or in heat conduction; see [Hi], [Ge], [SG], [FH], [SAWD], [SW], [Sh]) give rise to very interesting hyperbolic free boundary problems, which call for a thorough investigation. Problems of this kind also arise for systems of conservation laws, when existence and stability of multidimensional shock waves are addressed (see [Ma1], [Ma2], [Me]). In this connection, let us mention that quasi-linear and linear first-order hyperbolic problems with unknown boundaries are investigated in [LY], [Li], [KM].

In all the above references only *smooth* solutions are considered. On the other hand, it is well known that for first-order conservation laws we cannot expect to have global classical solutions (see [La], [Ol]). Hence in these cases a weak entropy formulation is unavoidable, both for Cauchy and for initial-boundary value problems (e.g., see [Kr], [BLN]).

To our knowledge, no *entropy formulation of free boundary problems* involving first-order conservation laws has been given so far. This is our concern in the present paper. More precisely, we consider the following Stefan problem:

$$\begin{cases} u_t + f(u)_x = 0 & \text{in } \Lambda(s(\cdot), T), \\ u(x,0) = u_0(x) & \text{in } \mathbb{R}^+ \times \{0\}, \\ s'(t) = K(u(s(t),t), s(t)), \\ s(0) = 0, \end{cases} \tag{1.1}$$

where f, u_0, K are given and u, s must be determined; here

$$\Lambda(s(\cdot), T) := \{(x,t) \in \mathbb{R} \times (0,T) \,:\, 0 \le t < T,\, s(t) < x\}\,.$$

To understand the nature of the problem some other comments are in order. As is well known, a typical feature of the entropy formulation of initial-boundary

*Received by the editors July 31, 1997; accepted for publication (in revised form) September 9, 1998; published electronically July 22, 1999.
http://www.siam.org/journals/sima/30-5/32530.html
†Dipartimento di Matematica "G. Castelnuovo," Universitá degli Studi "La Sapienza," Roma (terracin@mat.uniroma1.it).

value problems for scalar conservation laws is that the data are not strongly achieved; instead, boundary conditions have to be formulated as compatibility conditions in a suitable sense (see [BLN]). Roughly speaking, such compatibility conditions involve properties both of the flux function f and of the boundary (see Definition B.1 in Appendix B). Clearly, giving an entropy formulation for a free boundary problem like (1.1), where *the boundary is a priori unknown*, raises interesting, nontrivial questions.

Since the equation for the boundary $\{(s(t), t)\}$ in problem (1.1) involves the function K, we cannot expect this problem to be well defined unless the function has suitable properties. Investigating this point for a general Stefan problem like (1.1) appears to be a very difficult task. In this paper we limit ourselves to the following choice:

$$K(u, v) = \frac{f(u)}{u - g'(v)}\ , \tag{1.2}$$

which is suggested by a model for ion-etching proposed in [Ro] (see Appendix A). Here f is the flux and g is a regular decreasing function; in [Ro] the case was considered

$$g(x) \equiv -cx, \qquad c > 0, \tag{1.3}$$

with a constant initial datum $u_0(x) \equiv p$, $p > 0$.

Problem (1.1) with the function K given in (1.2) will be referred to as problem (FB) in the following. It is the purpose of this paper to investigate thoroughly this case, regarded as a "case study" toward the investigation of the general Stefan problem (1.1). As we shall see, some results concerning the general free boundary problem (1.1) can be proved using the same ideas (Proposition 2.4).

The main technical tool used in this paper is the comparison principle for mixed valued problems of scalar conservation laws which will be recalled in Appendix B.

2. The main results. In this section we introduce our mathematical framework and state the main results of the paper.

Concerning the flux f the following assumptions will be used:

(f_1) $f \in C^2(\mathbb{R})$;

(f_2) $f > 0$;

(f_3) $f(x) = f(-x)$ for any $x \in \mathbb{R}$;

(f_4) $f(x) \underset{|x|\to\infty}{\to} \longrightarrow 0$;

(f_5) f has a finite number of inflection points.

As for g, we assume that

(g) $g \in C^1(\mathbb{R}^+)$ is decreasing.

We also assume $u_0 \geq 0$.

Let us first state a local existence result concerning classical solutions of problem (1.1).

PROPOSITION 2.1. *Let $u_0 \in C^1(\mathbb{R}^+)$, $K \in C^1(\mathbb{R} \times \mathbb{R})$. Then there is $T > 0$ such that for $t \in (0, T)$ there exists at least one regular solution (s, u) of the free boundary problem* (1.1).

It was observed in [Ro] that even smooth solutions of problem (1.1) need not be unique. Hence the class of solutions we consider has to be restricted using a criterion of physical admissibility, which was introduced in [Ro] assuming (1.3) and constant initial data. Define

$$F_c(p) := \inf\left\{ q \in (p, \infty)\ :\ \frac{f(q)}{q + c} \geq \frac{f(r)}{r + c} \text{ for any } r \in (q, \infty) \right\}. \tag{2.1}$$

Then an admissible solution of problem (1.1)–(1.3) is given by a couple (s,u), such that $s(t)\equiv F_c(p)t$ and u solves the initial-boundary value problem:

$$\begin{cases} u_t+f(u)_x=0 & \text{in } \Lambda(s(\cdot),T),\\ u(x,0)=p & \text{in } \mathbb{R}^+\times\{0\},\\ u(F_c(p)t,t)=F_c(p) & \text{for every } t\in(0,T). \end{cases}$$

This characterizes the admissible solution as the one for which the speed of the interface is maximal. Observe that the function F_c is well defined, due to the properties of the flux function f. In some intervals this is equal to the identity; in other intervals it is constant. Since the flux f is decreasing for large x, the same is true for the function $\frac{f(p)}{p+c}$—hence the function F_c is equal to the identity for large x.

Similar ideas can be used to obtain the weak entropy formulation of problem (FB). We recall that in the standard space BV of functions with (locally) bounded total variation it is possible to define the notion of trace; see, for instance, [EG]. For simplicity we denote by $u(s(t),t)$ the trace (in the L^1 sense) of the boundary value function u along the (smooth) curve $s=s(t)$.

Let us introduce the following definition.

DEFINITION 2.1. *Let $u_0\in BV(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+)$. We say that the couple $s(t)\in Lip(0,T)$, $u\in BV_{loc}(\Lambda(s(\cdot),T))\cap L^\infty(\Lambda(s(\cdot),T))$, $u\geq 0$, is an entropy solution of problem* (FB) *if it is an entropy solution of the initial-boundary value problem*

$$\begin{cases} u_t+f(u)_x=0 & \text{in}\quad \Lambda(s(\cdot),T),\\ u(x,0)=u_0(x) & \text{in}\quad \mathbb{R}^+\times\{0\},\\ u(s(t),t)=F_{|g'(s(t))|}(u(s(t),t)),\\ s'(t)=\dfrac{f(F_{|g'(s(t))|}(u(s(t),t)))}{F_{|g'(s(t))|}(u(s(t),t))+|g'(s(t))|}. \end{cases}\tag{2.2}$$

The definition of entropy solution of an initial-boundary value problem for scalar conservation laws is recalled in Appendix B (see Definition B.1).

In particular, the boundary condition is satisfied in the following sense: For almost any $t\in(0,T)$ there holds

$$\frac{f(u(s(t),t))-f(k)}{u(s(t),t)-k}\leq\frac{f(F_{|g'(s(t))|}(u(s(t),t)))}{F_{|g'(s(t))|}(u(s(t),t))+|g'(s(t))|}$$

for any $k\in(u(s(t),t),F_{|g'(s(t))|}(u(s(t),t))]$.

Consider the set

$$\mathcal{S}(c):=\{p\in\mathbb{R}\,:\,F_c(p)=p\}.\tag{2.3}$$

The above set plays a crucial role in the entropy formulation of problem (FB), as the following proposition shows.

PROPOSITION 2.2. *Let $u_0\in BV(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+)$. The couple $s(t)\in Lip(0,T)$, $u\in BV_{loc}(\Lambda(s(\cdot),T))\cap L^\infty(\Lambda(s(\cdot),T))$ is an entropy solution of the problem* (FB) *if and only if this couple is an entropy solution of*

$$\begin{cases} u_t+f(u)_x=0 & \text{in}\quad \Lambda(s(\cdot),T),\\ u(x,0)=u_0(x) & \text{in}\quad \mathbb{R}^+\times\{0\},\\ u(s(t),t)\in S(|g'(s(t))|) & \text{almost everywhere (a.e.) in } (0,T),\\ s'(t)=\dfrac{f(u(s(t),t))}{u(s(t),t)+|g'(s(t))|}. \end{cases}\tag{2.4}$$

It follows that the trace $u(s(t),t)$ of any entropy solution of problem (FB) takes values in the set $\mathcal{S}(c)$. This implies that no characteristic starting from the boundary and pointing inwards toward the domain exists (Proposition 3.1 (v)).

The following proposition gives a maximality property of the entropy solutions of the free boundary problem in $\Lambda(s(\cdot),T)$.

PROPOSITION 2.3. *Let (s,u) be an entropy solution of problem* (FB). *Let v be any entropy solution of the conservation law in $\Lambda(s(\cdot),T)$, which satisfies the same initial condition as u. Then $v \le u$ a.e. in $\Lambda(s(\cdot),T)$.*

In the spirit of the previous proposition we can prove the following theorem, which gives an a priori estimate for the entropy solutions of problem (FB).

THEOREM 2.1. *Let $u_0 \in BV_{loc}(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+)$. Let the couple (s,u) be an entropy solution of the free boundary problem* (FB). *Then*

$$\text{ess inf}(u_0) \le u \le F_{J(T)}(\text{ess sup}\, u_0) \qquad \text{for a.e. } (x,t) \in \Lambda(s(\cdot),T), \tag{2.5}$$

where

$$J(T) := \sup_{t\in(0,T)} |g'(s(t))|.$$

Moreover, the function u is an entropy solution of the following boundary value problem:

$$\begin{cases} u_t + f(u)_x = 0 & \text{in} \quad \Lambda(s(\cdot),T), \\ u(x,0) = u_0(x) & \text{in} \quad \mathbb{R}^+ \times \{0\}, \\ u(s(t),t) = F_{J(T)}(\text{ess sup}\, u_0). \end{cases}$$

As discussed in section 1, it is natural to ask whether the previous results are still true for the general Stefan problem (1.1). In this connection let us make the following definition.

DEFINITION 2.2. *Let $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$. We say that the couple $s(t) \in Lip(0,T)$, $u \in BV_{loc}(\Lambda(s(\cdot),T)) \cap L^\infty(\Lambda(s(\cdot),T))$ is an entropy solution of problem* (1.1) *if*

(i) *the function u is an entropy solution in $\Lambda(s(\cdot),T)$ of problem* (1.1);

(ii) $s'(t) = K(u(s(t),t),s(t))$;

(iii) $u(s(t),t) \in \overline{\mathcal{S}}(s(t))$, *where the set $\overline{\mathcal{S}}(s(t))$ is a finite union of closed intervals and verifies the following property:*

$$u \in \overline{\mathcal{S}}(s(t)) \Leftrightarrow K(u,s(t)) \ge \frac{f(u)-f(p)}{u-p} \qquad \text{for every } u,p \in \overline{\mathcal{S}}(s(t)). \tag{2.6}$$

Following the proof of Proposition 2.3 (see section 3) we obtain the proposition below.

PROPOSITION 2.4. *Let (s,u) be an entropy solution of problem* (1.1). *Then the conclusion of Proposition* 2.3 *holds. Moreover,*

$$\sup\{q > \text{ess inf}\, u_0 \,:\, q \in \overline{\mathcal{S}}(s(t)) \ \forall t \in [0,T]\} \le u \le$$

$$\inf\{q > \text{ess sup}\, u_0 \,:\, q \in \overline{S}(s(t)) \ \forall t \in [0,T]\} \quad \text{a.e. in } \Lambda(s(\cdot),T).$$

Let us now restrict ourselves to the case $g(x) = -cx$, $c > 0$, which is easier since the admissible set of free boundary value $\mathcal{S}(|g'(s(t))|)$ does not depend on t. The

general case, namely, for any g satisfying assumption (g), will be the object of further investigation.

Let us introduce the following classes of initial data:

Class A: u_0 takes values in an interval $[a,b] \subset \mathcal{S}(c)$;

Class B: u_0 takes values in an interval $(p,q) \subset \mathbb{R}^+ \setminus \mathcal{S}(c)$;

Class C: u_0 has a finite number of oscillations outside a set of zero Lebesgue measure, namely, there exist a null set E and $n \in \mathbb{N}$ such that for any $n+1$ points $\{x_i\}$, $x_i < x_{i+1}$ $(i = 1, \dots, n)$ in $\mathbb{R}^+ \setminus E$, $u_0(x_j) \neq u_0(x_{(j+1)})$ there holds

$$sgn(u_0(x_{i+1}) - u_0(x_i)) = -sgn(u_0(x_{i+2}) - u_0(x_{i+1})), \qquad 1 \leq i \leq n-2,$$

implying

$$sgn(u_0(x_{n+1}) - u_0(x_n)) = sgn(u_0(x_n) - u_0(x_{n-1})).$$

Let us observe that $\mathcal{S}(c) = \mathbb{R}^+$ when $f' \leq 0$ (see Proposition 3.1 (i) below), in which case all the data belong to Class A.

The following existence and uniqueness theorem for problem (FB) can be proved (see sections 4–6).

THEOREM 2.2. *Let* (1.3) *hold. Let* $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ *be in any Class* A, B, *or* C. *Then there exists a unique entropy solution of the free boundary problem* (FB).

If the initial data are in Class A, the following estimates can be proved.

THEOREM 2.3. *Let* (1.3) *hold. Let* (s_1, u), (s_2, v) *be solutions of the free boundary problem* (FB) *with data* u_0, v_0, *respectively, of Class* A. *Then*

$$\|s_1 - s_2\|_{L^\infty(0,T)} \leq \frac{2}{c} \|u_0 - v_0\|_{L^1((0,R_T))},$$

and for any $C \in \mathbb{R}^+$ *and almost any* t

$$\int_{\overline{s}(t)}^{\overline{s}(t)+C} |u(x,t) - v(x,t)|\, dx \leq \|u_0 - v_0\|_{L^1((0,C+R_T))},$$

where

$$\overline{s}(t) = \max\{s_1(t), s_2(t)\},$$

$$R_T := \left(\sup_{|u| \leq M} |f^{'}(u)| + \|s'\|_{L^\infty(0,T)} \right) T,$$

and

$$M := \max(\|u_0\|_{L^\infty}, \|v_0\|_{L^\infty}).$$

Moreover, suppose $u_0 \leq v_0$ *a.e. Then*

$$s_1 \geq s_2 \qquad \text{in } (0,T),$$

$$u \leq v \qquad \text{a.e. in } \Lambda(s_1(t), T).$$

3. Proofs of general results. In this section we prove local existence of classical solutions of the general free boundary problem (1.1) (Proposition 2.1), as well as the characterization of the entropy solutions of the free boundary problem (FB) (Proposition 2.3) and the a priori estimates stated in Proposition 2.4 and Theorem 2.1, respectively.

Proof of Proposition 2.1. Let us extend u_0, in an arbitrary way, to a function $\overline{u}_0 \in C^1(\mathbb{R})$.

By the method of characteristics, there exists a time T_0 and a regular solution u to the Cauchy problem

$$\begin{cases} u_t + f(u)_x = 0 & \text{in} \quad \mathbb{R} \times (0, T_0), \\ u(x,0) = \overline{u}_0 & \text{in} \quad \mathbb{R} \times \{0\}. \end{cases}$$

We look for a solution $s(t)$ of

$$s'(t) = K(u(s(t),t), s(t)).$$

Set

$$B := \{s \in Lip(0,T) \,:\, s(0) = 0,\, 0 \le s \le M\,,\, 0 \le s' \le L \,\text{a.e.}\},$$

where $T < T_0$ is a constant to be fixed later,

$$L := \max_{\substack{|u| \le \|\overline{u}_0\|_{L^\infty(\mathbb{R})} \\ |v| \le 1}} |K(u,v)| \text{ and } M := \min\{LT, 1\}.$$

Observe that B is a closed subspace of the Banach space $C(0,T)$; consider the following operator G from B into itself:

$$G(s)(t) := \int_0^t K(u(s(t),t), s(t))\, d\tau.$$

Let us choose T such that G is a contraction. We have

$$\|G(s) - G(\overline{s})\|_{L^\infty} = \sup_{t \in [0,T]} \left| \int_0^t K(u(s(\tau),\tau), s(\tau)) - K(u(\overline{s}(\tau),\tau), \overline{s}(\tau))\, d\tau \right|.$$

Since K and u are smooth functions there exists a constant k such that

$$|K(u(x,t),x) - K(u(\overline{x},t),x)| \le k|x - \overline{x}|$$

for every $-M \le x, \overline{x} \le M$, $0 \le t \le T_0$.

Thus we have

$$\|G(s) - G(\overline{s})\|_{L^\infty} \le Tk \|s - \overline{s}\|_{L^\infty}.$$

Choosing $T < \frac{1}{k}$ we obtain that T is a contraction operator and we can apply the contraction fixed point theorem. □

Observe that the above result cannot give any information about uniqueness, since we consider an arbitrary extension of u_0. However, the following holds.

LEMMA 3.1. *Suppose $u_0 \in C^1(\mathbb{R}^+)$. Let u be a local regular solution given in the previous proof such that the trace $u(s(t),t)$ lies in the set $\overline{\mathcal{S}}(s(t))$ given by* (2.6)*. Then the function u does not depend on the arbitrary extension of u_0.*

Proof. Let us prove that $s(t)$ depends only on the data on the positive semiaxis.

For any $x_0 \geq 0$ the characteristic issued from x_0 is given by the solution of the problem

$$\begin{cases} x' = f(u), \quad u' = 0, \\ x(0) = x_0, \quad u(0) = u(x(0)) = u(x_0). \end{cases}$$

Then we have

$$x(t) = x_0 + tf'(u_0(x_0)), \quad u(x(t), t) = u_0(x_0).$$

Let $(x(t), 0)$ be the starting point of the characteristic reaching the point $(s(t), t)$; then

$$s(t) = x_0(t) + tf'(u_0(x_0(t))) \;\; s(0) = 0 = x_0(0). \tag{3.1}$$

Let us derive (3.1) obtaining

$$s'(t) = x_0'(t) + f'(u_0(x_0(t))) + tf''(u_0(x_0(t)))x_0'(t)u_0'(x_0(t)).$$

Since $s'(t) = K(u(s(t), t), s(t))$, we obtain by (2.6)

$$x_0'(1 + tf''u_0') = s'(t) - f'(u_0(x_0(t))) = K(u(s(t), t), t) - f'(u(s(t), t)) \geq 0.$$

Thus for small time $x_0' \geq 0$ and $x_0(t) \geq 0$. □

Let us now state some properties of the function F_c and of the set $S(c)$ (see (2.1), (2.3)), which are needed in the following. The elementary proofs are omitted.

PROPOSITION 3.1. *Assume* (f_1)–(f_5) *for the flux function* f. *Then*

(i) $F_c(p) \geq p$ *for any* $p \geq 0$;

(ii) *let* $F_c(p) > p$. *Then* $\frac{f(F_c(p))}{F_c(p)+c} = f'(F_c(p))$ *and* $\frac{f(F_c(p))}{F_c(p)+c} > \frac{f(p)}{p+c}$;

(iii) *there exists a finite number of points* n, $0 = p_0 \leq p_1 < p_2 < \cdots < p_n = +\infty$, *such that*

$$F(p) = \begin{cases} p & \text{in } (p_{2j}, p_{2j+1}], \\ F(p_{2j+2}) & \text{in } (p_{2j+1}, p_{2j+2}]; \end{cases}$$

(iii) F *is continuous from the left;*

(iv) *let* $u \in \mathcal{S}(c)$, $u < v$. *Then*

$$\min\left\{\frac{f(u)}{u+c}, \frac{f(v)}{v+c}\right\} \geq \frac{f(u) - f(v)}{u - v};$$

(v) *for any given interval* $(a, b) \subset \mathcal{S}(c)$, *the function* $\frac{f(x)}{x+c}$ *is nonincreasing in* (a, b) *and*

$$\frac{f(x)}{x + c} - f'(x) \geq 0.$$

Let us discuss a more precise characterization of the set $\mathcal{S}(c)$. For this purpose, let $0 < p_1 < \cdots < p_k$ local strictly positive maximum of f'. Define the functions

$$L(x) = x - \frac{f(x)}{f'(x)},$$

$$p_j(c) := \begin{cases} \min\{x \geq p_j : L(c) = -c,\ f''(x) \geq 0\}, & j = 1, \ldots, k, \\ 0, & j = 0. \end{cases}$$

Set also

$$N(c) := \{j \in \{0, \dots, k\} \,:\, p_j(c) \in \mathcal{S}(c)\},$$

and for any $j \in K(c) \setminus \{0\}$

$$\overline{p}_j(c) := \max\left\{p_{j-1}(c) \le x < p_j(c) : \frac{f(x)}{x+c} = f'(p_j(c))\right\}.$$

The following result can be proved.

PROPOSITION 3.2. (i) *If $f' \le 0$, then $\mathcal{S}(c) = \mathbb{R}^+$ for every c.*

(ii) *The function $p_j(c)$ is an increasing function of c for every $j = 1, \dots, k$.*

(iii) *The function $\overline{p}_j(c)$ is a decreasing function of c for every $j \in N(c) \setminus \{0\}$.*

(iv) *There holds*

$$\mathcal{S}(c) = \bigcup_{j \in N(c) \setminus \{\max\{k \in N(c)\}\}} [p_j(c), \overline{p}(c)_{\min\{k \in N(c),\, k > j\}}] \cup [p(c)_{\max\{k \in N(c)\}}, \infty).$$

(v) *$\mathcal{S}(c_1) \subset \mathcal{S}(c_2)$ for any $c_1 > c_2$.*

We can now prove Propositions 2.2 and 2.3.

Proof of Proposition 2.2. If u is an entropy solution of (2.4), then it is obviously an entropy solution of (2.2). Now suppose that the function u is an entropy solution of (2.2). We want to show that the trace $u(s(t), t) \in S(|g'(s(t))|)$ a.e.

For simplicity set $u(s(t), t) = \tilde{u}$ for a fixed t. If $\tilde{u} < F_{|g'(s(t))|}(\tilde{u})$ then for the compatible boundary conditions we have

$$\frac{f(\tilde{u}) - f(k)}{\tilde{u} - k} \le s'(t) = \frac{f(F_{|g'(s(t))|}(\tilde{u}))}{F_{|g'(s(t))|}(\tilde{u}) + |g'(s(t))|} \qquad \text{for every } k \in (\tilde{u}, F_{|g'(s(t))|}(\tilde{u})].$$

Taking $k = F_{|g'(s(t))|}(\tilde{u})$, we obtain by computations

$$\frac{f(F_{|g'(s(t))|}(\tilde{u}))}{F_{|g'(s(t))|}(\tilde{u}) + |g'(s(t))|} \le \frac{f(\tilde{u})}{\tilde{u} + |g'(s(t))|}.$$

This is in contradiction with Proposition 3.1 (iv). □

Let us associate with any given function $s \in Lip(0, T)$ and $u \in BV(\Lambda(s(\cdot), T)) \cap L^\infty(\Lambda(s(\cdot), T))$, the function

$$u^s(x, t) := u(x + s(t), t) \tag{3.2}$$

defined in $\Pi_T^+ := \mathbb{R}^+ \times (0, T)$.

Proof of Proposition 2.3. We have to compare the functions u and v in $\Lambda(s(\cdot), T)$. Consider the functions u^s, v^s associate, respectively, to $(s(t), u)$ and $(s(t), v)$ (see (3.2)). Using the inequality (B.4) of Theorem B.1 (see Appendix B), we obtain

$$\begin{aligned} &\int_{\mathbb{R}^+} [v^s(x,t) - u^s(x,t)]_+ \, dx \\ &\qquad \le \int_0^t H(v^s(0,\tau) - u^s(0,\tau))(f(v^s(0,\tau)) - f(u^s(0,\tau)) \\ &\qquad - \frac{f(u^s(0,\tau))}{u^s(0,\tau) + |g'(s(t))|}(v^s(0,\tau) - u^s(0,\tau))) \, d\tau. \end{aligned} \tag{3.3}$$

Let us show that the right-hand side of inequality (3.3) is negative. In fact, for the values of τ such that $v^s(0,\tau) \le u^s(0,\tau)$ the integrand is negative. Suppose $v^s(0,\tau) > u^s(0,\tau)$. Since $u^s \in \mathcal{S}(|g'(s(\tau))|)$, we have from Proposition 3.1 (iv)

$$\frac{f(v^s(0,\tau)) - f(u^s(0,\tau))}{v^s(0,\tau) - u^s(0,\tau)} \le \frac{f(u^s(0,\tau))}{u^s(0,\tau) + |g'(s(\tau))|}.$$

Therefore also in this case the integrand is negative. This completes the proof. ∎

Let us prove Theorem 2.1.

Proof of Theorem 2.1. Set $p := F_{J(T)}(\operatorname{ess\,sup} u_0)$. Consider in $\Lambda(s(\cdot),T)$ the functions u and $v \equiv p$ and the associated functions u^s and v^s given by (3.2). Since $v^s \equiv p$ is an entropy solution of the problem with initial data $v_0 \equiv p$, we can compare the functions u^s and v^s by inequality (B.4). In this case we have

$$\int_{\mathbb{R}^+} [u^s(x,t) - p]_+ \, dx \le \int_{\mathbb{R}^+} [u(x,0) - p]_+ \, dx$$

$$+ \int_0^t H(u^s(0,\tau) - p) \left(f(u^s(0,\tau)) - f(p) - \frac{f(u^s(0,\tau))}{u^s(0,\tau) + c}(u^s(0,\tau) - p) \right) d\tau. \tag{3.4}$$

Again using Proposition 3.1 (iv) and Proposition 3.2 (v) we obtain that the right-hand side of inequality (3.4) is negative. In the same way we can prove that $u^s \ge \operatorname{ess\,inf} u_0$.

The rest of the proof is a consequence of Theorem B.2 (see Appendix B), Proposition 2.3, and estimate (2.5). ∎

Let us prove for future reference a further result. Set

$$M_0 = \max\{\|u_0\|_{L^\infty}, F_{J(T)}(\operatorname{ess\,sup} u_0)\},$$

where

$$J(T) := \sup_{t \in (0,T)} |g'(s(t))|.$$

Then the following holds.

PROPOSITION 3.3. *Let $u_0 \in BV_{loc}(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$. Let (s,u) be an entropy solution of the free boundary problem* (FB). *Let $\overline{u}_0 \in BV_{loc}(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ such that $\|\overline{u}_0\|_{L^\infty} \le \|u_0\|_{L^\infty}$ and*

$$\overline{u}_0 \equiv u_0 \qquad \text{in } [0,R]$$

for some $R > (\sup_{|u| \le M_0} |f'(u)| + \|s'\|_{L^\infty(0,T)})T$. Then a solution of the free boundary problem with initial data $\overline{u}_0$ is given by $(s,\overline{u})$, where $\overline{u}$ is the solution of the following mixed value problem:

$$\begin{cases} \overline{u}_t + f(\overline{u})_x = 0 & \text{in} \quad \Lambda(s(\cdot),T), \\ \overline{u}(x,0) = \overline{u}_0(x) & \text{in} \quad \mathbb{R}^+ \times \{0\}, \\ \overline{u}(s(t),t) = F_{J(T)}(\operatorname{ess\,sup} u_0). \end{cases}$$

Proof. Using Proposition B.1, Theorem B.2, Proposition 2.3, and Theorem 2.1 we easily obtain $\overline{u}(s(t),t) = u(s(t),t)$ for almost any t in $(0,T)$. ∎

4. Initial data of Class A. In this section we prove Theorem 2.2 for initial data u_0 in the Class A, as well as Theorem 2.3. Observe that by Theorem 2.1 any entropy solution of problem (FB) takes values in the interval $[a,b] \subset \mathcal{S}(c)$.

We first prove the uniqueness claim of Theorem 2.2, as well as stability and comparison results which imply plainly Theorem 2.3 (see subsection (a)). Then the existence part of Theorem 2.2 is proved by a suitable adaptation of the Godunov method (see subsection (b)).

(a) *Uniqueness and comparison results.*

PROPOSITION 4.1. *Let* (1.3) *hold; let* $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ *belong to Class* A. *Then there is at most one entropy solution of problem* (FB).

Proof. Let (s_1,u_1), (s_2,u_2) be two entropy solutions of problem (FB). Let $\overline{s}(t) = \max\{s_1(t), s_2(t)\}$. Obviously $\overline{s}(t) \in Lip(0,T)$.

Consider the associate functions ${u_1}^{\overline{s}}$, ${u_2}^{\overline{s}}$ relative to u_1, u_2 with respect to the function $\overline{s}$ (see (3.2)). Then by Proposition B.1, functions ${u_1}^{\overline{s}}$, ${u_2}^{\overline{s}}$ are both entropy solutions of the problem

$$\begin{cases} u_t + f(u)_y - \overline{s}'(t)u_y = 0 & \text{in} \quad \Pi_T^+, \\ u(y,0) = u_0(y) & \text{in} \quad \mathbb{R}^+ \times \{0\}. \end{cases}$$

Now, using Theorem B.1, we have for almost any $t \in (0,T)$

$$\int_{\mathbb{R}^+} |{u_1}^{\overline{s}}(x,t) - {u_2}^{\overline{s}}(x,t)|\,dx \tag{4.1}$$

$$\leq \int_0^t sgn({u_1}^{\overline{s}}(0,\tau) - {u_2}^{\overline{s}}(0,\tau))(f({u_1}^{\overline{s}}(0,\tau))$$

$$- f({u_2}^{\overline{s}}(0,\tau)) - \overline{s}'(\tau)({u_1}^{\overline{s}}(0,\tau) - {u_2}^{\overline{s}}(0,\tau)))\,d\tau.$$

Let us show that the right-hand side in inequality (4.1) is negative.

For almost any $t \in (0,T)$ there is $i \in \{1,2\}$ such that $\overline{s}'(t) = \frac{f({u_i}^{\overline{s}}(0,t))}{{u_i}^{\overline{s}}(0,t)+c}$. Fix a t in the set in which the previous equality holds.

The integrand of the right-hand side of inequality (4.1) can be written

$$|{u_1}^{\overline{s}}(0,t) - {u_2}^{\overline{s}}(0,t)| \left(\frac{f({u_1}^{\overline{s}}(0,t)) - f({u_2}^{\overline{s}}(0,t))}{{u_1}^{\overline{s}}(0,t) - {u_2}^{\overline{s}}(0,t)} - \overline{s}'(t) \right). \tag{4.2}$$

Since ${u_1}^{\overline{s}}(0,t)$, ${u_2}^{\overline{s}}(0,t)$ lay in $\mathcal{S}(c)$, we have that (4.2) is negative by Proposition 3.1 (iv). Then ${u_1}^{\overline{s}}$ and ${u_2}^{\overline{s}}$ coincide. Therefore $u_1 \equiv u_2$ in $\Lambda(\overline{s}(t),T)$.

Hence $(\overline{s}, u_1(\equiv u_2))$ is another entropy solution of problem (FB).

Let us show that such a solution is equal to (s_i,u_i) $(i=1,2)$. To obtain this, it is enough to prove that $s_1 \equiv s_2 \equiv \overline{s}$. By Proposition 3.3 it is not restrictive to assume $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+)$.

Using conservation formula (B.6) in the cases $s(t) = s_i(t)$ and $s(t) = \overline{s}(t)$, we have

$$\int_{s_1(t)}^\infty u_1(x,t)\,dx = \int_{\mathbb{R}^+} u_0(x)\,dx + cs_1(t) - tf(0)$$

and

$$\int_{\overline{s}(t)}^\infty u_1(x,t)\,dt = \int_{\mathbb{R}^+} u_0(x)\,dx + c\overline{s}(t) - tf(0).$$

Subtracting the previous equalities, we obtain

$$\int_{s_1(t)}^{\infty} u_1(x,t)\,dx - \int_{\overline{s}(t)}^{\infty} u_1(x,t)\,dx = c(s_1(t) - \overline{s}(t)).$$

Hence

$$0 \geq c(s_1(t) - \overline{s}(t)) = \int_{s_1(t)}^{\overline{s}(t)} u_1(x,t)\,dx \geq 0,$$

which implies $\overline{s} \equiv s_1$. The same is true for $i = 2$ and so $s_2 \equiv s_1$ and $u_2 = u_1$. □

By using a similar argument we can prove the following stability result. Let us set, for any u_0, $v_0 \in BV_{loc}(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$,

$$M := \max(\|u_0\|_{L^\infty}, \|v_0\|_{L^\infty})$$

and

$$R_T := \left(\sup_{|u| \leq M} |f^{'}(u)| + \|s'\|_{L^\infty(0,T)} \right) T.$$

PROPOSITION 4.2. *Let* (1.3) *hold; let* u_0, $v_0 \in BV_{loc}(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ *and belong to Class* A. *Let* (s_1, u), (s_2, v) *be solutions of problem* (FB) *with data* u_0, v_0, *respectively. Then*

$$\|s_1 - s_2\|_{L^\infty(0,T)} \leq \frac{2}{c} \|u_0 - v_0\|_{L^1((0,R_T))}.$$

Moreover, for any $C \in \mathbb{R}^+$ *we have for almost any* t

$$\int_{\overline{s}(t)}^{\overline{s}(t)+C} |u(x,t) - v(x,t)|\,dx \leq \|u_0 - v_0\|_{L^1((0,C+R_T))},$$

where $\overline{s}(t) = \max\{s_1(t), s_2(t)\}$.

Proof. Proceeding as in the proof of Proposition 4.1 and using Theorem B.1 we obtain

$$\int_{\overline{s}(t)}^{\overline{s}(t)+C} |u(x,t) - v(x,t)|\,dx \leq \|u_0 - v_0\|_{L^1((0,C+R))}.$$

Let $\overline{u}_0 \equiv \chi_{(0,R_T)} u_0$, $\overline{v}_0 \equiv \chi_{(0,R_T)} v_0$, where $\chi_{(0,R_T)}$ is the characteristic function of the interval $(0, R_T)$. Then by Proposition 3.3 we see that there exist $(s_1, \overline{u})$, $(s_2, \overline{v})$, respectively, solutions of the free boundary with initial data $\overline{u}_0$, $\overline{v}_0$, respectively.

Obviously for the functions $\overline{u}$, $\overline{v}$ the previous inequality is still true. Moreover we have the conservation equalities

$$\int_{s_1(t)}^{\infty} \overline{u}(x,t)\,dx = \int_{\mathbb{R}^+} \overline{u}_0(x)\,dx + cs_1(t) - tf(0),$$

$$\int_{s_2(t)}^{\infty} \overline{v}(x,t)\,dx = \int_{\mathbb{R}^+} \overline{v}_0(x)\,dx + cs_2(t) - tf(0).$$

Fix t and suppose, for example, that $s_1(t) > s_2(t)$. Subtracting the previous equalities we have

$$c(s_1(t) - s_2(t))$$

$$= \int_{s_1(t)}^{\infty} \overline{u}(x,t) - \overline{v}(x,t)\,dx - \int_{s_2(t)}^{s_1(t)} \overline{v}(x,t)\,dx$$

$$+ \int_{\mathbb{R}^+} \overline{v}_0(x) - \overline{u}_0(x)\,dx \le 2\|\overline{u}_0 - \overline{v}_0\|_{L^1(\mathbb{R}^+)} = 2\|u_0 - v_0\|_{L^1((0,R))}.$$

From the arbitrariness of t we obtain the assert. □

Finally, let us state a comparison result.

PROPOSITION 4.3. *Let* (1.3) *hold; let* $u_0, v_0 \in BV_{loc}(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ *belong to Class* A *with* $u_0 \le v_0$ *a.e. Then* $s_1 \ge s_2$ *and* $u \le v$ *a.e. in* $\Lambda(s_1(t), T)$, *where* (s_1, u), (s_2, v) *are the solutions of the free boundary problem* (FB) *with data* u_0, v_0, *respectively.*

Proof. Let $\overline{s}(t) = \max\{s_1(t), s_2(t)\}$. As was shown previously we compare u and v in $\Lambda(\overline{s}(t), T)$.

Using inequality (B.4) and proceeding as in the proof of Proposition 2.3 we conclude that $u \le v$ a.e. in $\Lambda(\overline{s}(t), T)$.

It remains to prove that $s_1 \ge s_2$. As in the proof of Proposition 4.1 we can reduce ourselves to the case $u_0, v_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+)$.

Let $t \in (0, T)$ and suppose $s_2(t) > s_1(t)$. Considering the conservation equalities for u, v we have

$$c(s_2(t) - s_1(t)) + \int_{\mathbb{R}^+} v_0(x) - u_0(x)\,dx$$

$$= \int_{s_2(t)}^{\infty} v(x,t) - u(x,t)\,dx - \int_{s_1(t)}^{s_2(t)} u(x,t)\,dx \le \int_{\mathbb{R}^+} v_0(x) - u_0(x)\,dx.$$

Hence we conclude that $s_2(t) \le s_1(t)$, which is a contradiction. □

(b) *Existence.*

Let us introduce the following standard notations. Let V_n be the space of the functions which are constant on any interval of the type

$$\left[\frac{i}{n}, \frac{i+1}{n}\right) \qquad (i \in \mathbb{N}).$$

Define for any $n \in \mathbb{N}$

$$P_n : BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+) \to V_n,$$

$$P_n(u)(t) := n \int_{\frac{i}{n}}^{\frac{i+1}{n}} u(\tau)\,d\tau, \qquad t \in \left[\frac{i}{n}, \frac{i+1}{n}\right).$$

The following lemma gives some interesting properties of P_n (see [Le1], [Le2], [Te2]).

LEMMA 4.1. *For any* $n \in \mathbb{N}$, $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+)$, *the operator* P_n *has the following properties:*

(i) $\|P_n(u_0)\|_{L^\infty(\mathbb{R}^+)} \le \|u_0\|_{L^\infty(\mathbb{R}^+)}$;
(ii) $TV(P_n(u_0), \mathbb{R}^+) \le TV(u_0, \mathbb{R}^+)$;
(iii) $\|P_n(u_0)\|_{L^1(\mathbb{R}^+)} \le \|u_0\|_{L^1(\mathbb{R}^+)}$;
(iv) $\|P_n(u_0) - u_0\|_{L^1(\mathbb{R}^+)} \le \frac{1}{2n} TV(u_0, \mathbb{R}^+)$.

Let us prove the following proposition.

PROPOSITION 4.4. *Let* (1.3) *hold; let* $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ *be initial data of Class* A. *Then there exists an entropy solution of problem* (FB).

Proof. It is not restrictive to assume $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+)$.

By Proposition 2.2 it is enough to look for an entropy solution of the problem

$$\text{(4.3)} \qquad \begin{cases} u_t + f(u)_x - \frac{f(u(0,t))}{u(0,t)+c} u_x = 0 & \text{in } \Pi_T^+, \\ u(0,t) \in \mathcal{S}(c) & \text{a.e.}, \\ u(x,0) = u_0(x) & \text{in } \mathbb{R}^+ \times \{0\}. \end{cases}$$

Namely, we search $u \in BV(\Pi_T^+) \cap L^\infty(\Pi_T^+)$ such that for any $\phi \in C_0^1((0,\infty) \times (0,T))$, $\phi \geq 0$, for any $k \in \mathbb{R}$,

$$\text{(4.4)} \qquad \int_{\Pi_T^+} |u-k|\phi_t + sgn(u-k)(f(u)-f(k))\phi_x - |u-k|\frac{f(u(0,t))}{u(0,t)+c}\phi_x \, dtdx \geq 0;$$

$$\operatorname{ess\,lim}_{t\to 0} \int_I |u(x,t) - u_0(x)| \, dx = 0.$$

Let us consider the following scheme.

Let $u_{0n} = P_n u_0$. Let

$$l := \sup_{p\in\mathbb{R}^+} |f'(p)| + \sup_{p\in\mathbb{R}^+} \left|\frac{f(p)}{p+c}\right|;$$

$$r := \frac{1}{2l}, \ \Delta t = \frac{r}{n};$$

we denote with u_i^0 the constant value of the function $P_n(u_0)$ on the interval $(\frac{i}{n}, \frac{i+1}{n})$.

Let us solve the following problem until time $t_1 = \Delta t$:

$$\begin{cases} u_t + f(u)_x - \frac{f(u_0^0)}{u_0^0+c} u_x = 0 & \text{in } \mathbb{R}^+ \times (0,t_1), \\ u(x,0) = P_n u_0 & \text{in } \mathbb{R}^+ \times \{0\}, \\ u(0,t) = u_0^0 & \text{in } \{0\} \times (0,t_1). \end{cases}$$

Denote the solution of such problems with u_{1n}. For the given choice of t_1, we observe that the solution is obtained by solving Riemann problems between the constant values. Such problems are not interacting within the definition of Δt.

At time t_1 we consider a new boundary value problem with initial data $P_n(u_{1n}(x,t_1))$. More precisely we consider the following problem:

$$\begin{cases} u_t + f(u)_x - \frac{f(u_0^1)}{u_0^1+c} u_x = 0 & \text{in } \mathbb{R}^+ \times (t_1,t_2), \\ u(x,0) = P_n(u_{1n}(x,t_1)) & \text{in } \mathbb{R}^+ \times \{t_1\}, \\ u(0,t) = u_0^1 & \text{in } \{0\} \times (t_1,t_2), \end{cases}$$

where u_0^1 is the value of $P_n(u_{1n}(x,t_1))$ in the interval $(0, \frac{1}{n})$ and $t_2 = 2t_1$. We call the solution of such problem u_{2n}.

By induction we can define a function u_n until time T.

Set

$$u_i^j := P_n(u_{(j-1)n}(\cdot, t_j)) \qquad \text{in the interval } \left(\frac{i}{n}, \frac{i+1}{n}\right).$$

Moreover we denote with $u^j_{i+\frac{1}{2}}$ the constant value of u_{jn} in the points $(\frac{i+1}{n}, t)$ (the dependence on n is understood in the previous notations).

Finally define

$$u_{n|_{\mathbb{R}^+\times(t_j,t_{j+1})}} := u_{jn}, \qquad j = 0, 1, \dots.$$

We obtain the proof in three steps:

(i) there exists a subsequence of $\{u_n\}$, still denoted by $\{u_n\}$, converging in the L^1 norm to a function $u \in BV(\Pi^+_T) \cap C([0,T], L^1_{loc}(\mathbb{R}^+)) \cap L^\infty(\Pi^+_T)$;

(ii) $u_n(0,\cdot) \in BV(0,T)$, with $\|u_n(0,\cdot)\|_{BV(0,T)} \le C$, where C is a constant independent of n;

(iii) the function u is an entropy solution of the free boundary problem (FB).

To prove (i), let us consider the total variation of $u_{jn}(\cdot, t_j)$,

$$TV(u_{jn}(\cdot,t_j), \mathbb{R}^+) = \sum_{i=0}^{\infty} |u^j_{i+1} - u^j_i|.$$

Since the function $u_{(j-1)n}$ is given by solving the Riemann problem we have the following estimate by standard results on the Godunov scheme (see [Le1], [Te2]):

$$\sum_{i=0}^{\infty} |u^j_{i+1} - u^j_i| \le \sum_{i=0}^{\infty} |u^{j-1}_{i+1} - u^{j-1}_i| \le TV(u_0)$$

for any j.

Then, for any t in $(0,T)$, there exists j such that $t_j \le t \le t_{j+1}$ and

$$TV(u_n(\cdot,t)) = TV(u_{jn}(\cdot,t)) \le TV(u_{jn}(\cdot,t_j)) \le TV(u_0) \tag{4.5}$$

obtaining an estimate independent of n.

In the same way we can prove the following estimate:

$$\sum_{i=0}^{\infty} |u^{j+1}_i - u^j_i| \le \sum_{i=0}^{\infty} |u^j_{i+1} - u^j_i| \le TV(u_0). \tag{4.6}$$

Let t, $s \in (0,T)$. Let us estimate $\|u_n(\cdot,t) - u_n(\cdot,s)\|_{L^1(\mathbb{R}^+)}$.

Suppose $t \in (t_j, t_{j+1})$, $s \in (t_l, t_{l+1})$, $l > j$. Then by (4.6)

$$\begin{aligned}
\|u_n(\cdot,t) - u_n(\cdot,s)\|_{L^1(\mathbb{R}^+)} &\le \|u_n(\cdot,t) - u_n(\cdot,t_{j+1})\|_{L^1(\mathbb{R}^+)} \\
&+ \sum_{k=j+1}^{l-1} \|u_n(\cdot,t_{k+1}) - u_n(\cdot,t_k)\|_{L^1(\mathbb{R}^+)} + \|u_n(\cdot,t_l) - u_n(\cdot,s)\|_{L^1(\mathbb{R}^+)} \\
&\le \|u_n(\cdot,t) - u_{jn}(\cdot,t_{j+1})\|_{L^1(\mathbb{R}^+)} + \|u_{jn}(\cdot,t_{j+1}) - u_n(\cdot,t_{j+1})\|_{L^1(\mathbb{R}^+)} \\
&+ \sum_{k=j+1}^{l-1} \sum_{i=0}^{\infty} \int_{\frac{i}{n}}^{\frac{i+1}{n}} |u^k_i - u^{k+1}_i|\, dx + \|u_n(\cdot,t_l) - u_n(\cdot,s)\|_{L^1(\mathbb{R}^+)} \\
&\le \|u_n(\cdot,t) - u_{jn}(\cdot,t_{j+1})\|_{L^1(\mathbb{R}^+)} + \frac{1}{2n} TV(u_0) \\
&+ \frac{(l-j-1)r}{rn} TV(u_0) + \|u_n(\cdot,t_l) - u_n(\cdot,s)\|_{L^1(\mathbb{R}^+)}.
\end{aligned}$$

Since u_{jn}, u_{ln} are piecewise smooth solutions, respectively, of the conservation laws

$$u_t + f(u)_x - \frac{f(u_0^j)}{u_0^j + c} = 0 \quad \text{in} \quad \mathbb{R}^+ \times (t_j, t_{j+1}),$$
$$u_t + f(u)_x - \frac{f(u_0^l)}{u_0^l + c} = 0 \quad \text{in} \quad \mathbb{R}^+ \times (t_l, t_{l+1}),$$

we see that the first and the fourth addendum in the last term of the previous inequality can be estimated by $\frac{(t_{j+1}-t)}{r} TV(u_0)$, respectively, $\frac{(s-t_l)}{r} TV(u_0)$ (see [Le2]). Therefore we have

$$\begin{aligned} &\|u_n(\cdot,t) - u_n(\cdot,s)\|_{L^1(\mathbb{R}^+)} \\ &\le \frac{(t_{j+1}-t)}{r} TV(u_0) + \frac{(s-t_l)}{r} TV(u_0) + \frac{(l-j-1)r}{rn} TV(u_0) + \frac{1}{2n} TV(u_0) \\ &= \frac{TV(u_0)}{r} \left(t_{j+1} - t + s - t_l + (l-j-1)\frac{r}{n} \right) TV(u_0) + \frac{1}{2n} TV(u_0) \\ &= (s-t)\frac{TV(u_0)}{r} + \frac{1}{2n} TV(u_0). \end{aligned} \tag{4.7}$$

The estimates (4.5)–(4.7) ensure us that step (i) holds. Moreover by the previous estimates we see that $u_n \in BV(\Pi_T^+)$ for any n. Observe that it is not restrictive to suppose that the sequence converges a.e. in Π_T^+ to u.

Let us consider assertion (ii). Fix n and consider $u_n(0,\cdot)$; then

$$TV(u_n(0,\cdot)) = \sum_{i=1}^{k(n)} |u_0^i - u_0^{i-1}|,$$

where $k(n)$ is the last step to reach T.

Let us prove the following estimate. For every l, $s \in \mathbb{N}$

$$\sum_{k=1}^{s} |u_k^l - u_{k-1}^l| + |u_0^l - u_0^{l-1}| \le \sum_{k=1}^{s+1} |u_k^{l-1} - u_{k-1}^{l-1}|. \tag{4.8}$$

Consider the following Cauchy problem:

$$\begin{cases} u_t + f(u)_x - \frac{f(u_0^{l-1})}{u_0^{l-1}+c} u_x = 0 \qquad \text{in} \quad \mathbb{R}^+ \times (t_{l-1}, t_l), \\ u(x, t_{l-1}) = \begin{cases} u_0^{l-1}, & x \le 0, \\ u_n(x, t_{l-1}), & x > 0, \end{cases} \qquad \text{in} \quad \mathbb{R} \times \{t_{l-1}\}. \end{cases}$$

We know that

$$TV(u(\cdot, t_{l-1}, \mathbb{R})) = \sum_{k=1}^{\infty} |u_k^{l-1} - u_{k-1}^{l-1}|;$$

moreover, by standard arguments (see, for instance, [Df]), we have

$$TV\left(u(\cdot,t_l),\, \left(-\infty, \frac{s}{n}\right)\right) \le TV\left(u(\cdot,t_{l-1}),\, \left(-\infty, \frac{s+1}{n}\right)\right).$$

Let us observe that $u(\cdot,t_l) = u_0^{l-1}$ when $x < 0$.

Moreover

$$\sum_{k=1}^{s} |u_k^l - u_{k-1}^l| + |u_0^l - u_0^{l-1}|$$

$$= TV\left(P_n(u(\cdot,t_l)),\left(-\infty,\frac{s}{n}\right)\right) \le TV\left(u(\cdot,t_l),\left(-\infty,\frac{s}{n}\right)\right)$$

$$\le TV\left(u(\cdot,t_{l-1}),\left(-\infty,\frac{s+1}{n}\right)\right) = \sum_{k=1}^{s+1} |u_k^{l-1} - u_{k-1}^{l-1}|$$

and then inequality (4.8) holds.

Let us show now that for any $s \in \mathbb{N},\ 0 \le s \le k(n)$,

$$\sum_{k=k(n)-s}^{k(n)} |u_0^k - u_0^{k-1}| \le \sum_{m=1}^{s+1} |u_m^{k(n)-s-1} - u_{m-1}^{k(n)-s-1}|. \tag{4.9}$$

We prove inequality (4.9) by induction.

By the explicit solution of Riemann problem we have

$$|u_0^{k(n)} - u_0^{k(n)-1}| \le |u_0^{k(n)-1} - u_1^{k(n)-1}|,$$

which corresponds to the case $s = 0$. Suppose that inequality (4.9) is true for $s-1$; then we have

$$\sum_{k(n)-s}^{k(n)} |u_0^k - u_0^{k-1}| = \sum_{k(n)-s+1}^{k(n)} |u_0^k - u_0^{k-1}| + |u_0^{k(n)-s} - u_0^{k(n)-s-1}|$$

$$\le \sum_{m=1}^{s} |u_m^{k(n)-s} - u_{m-1}^{k(n)-s}| + |u_0^{k(n)-s} - u_0^{k(n)-s-1}|.$$

Now applying inequality (4.8) with $l = k(n) - s$, we see that

$$\sum_{k(n)-s}^{k(n)} |u_0^k - u_0^{k-1}| \le \sum_{m=1}^{s+1} |u_m^{k(n)-s-1} - u_{m-1}^{k(n)-s-1}|$$

and (4.9) holds.

Putting $s = k(n) - 1$ in (4.9) we obtain

$$\sum_{k=1}^{k(n)} |u_0^k - u_0^{k-1}| \le \sum_{m=1}^{k(n)} |u_m^0 - u_{m-1}^0| \le TV(u_0).$$

Since we estimate $TV(u_n(0,\cdot),(0,T))$ independently on n, we see that there exists a subsequence that converges in L^1 norm and a.e. to a function $w \in L^1((0,T)) \cap L^\infty((0,T))$ and $w(t) \in \mathcal{S}(c)$ for almost any t.

Let us consider step (iii). We have to prove that u verifies (4.4) for any nonnegative $\phi \in C_0^1((0,\infty)\times(0,T))$ and any $k \in \mathbb{R}$.

For any n, j, $\phi \in C_0^1([0,\infty)\times(0,T))$ we obtain

$$\int_{\mathbb{R}^+}\int_{t_j}^{t_{j+1}} |u_{jn}-k|\phi_t + sgn(u_{jn}-k)(f(u_{jn})-f(k))\phi_x - \frac{f(u_0^j)}{u_0^j+c}|u_{jn}-k|\phi_x\,dtdx$$

$$+\int_{t_j}^{t_{j+1}}\left(sgn(u_0^j-k)(f(u_0^j)-f(k)) - \frac{f(u_0^j)}{u_0^j+c}|u_0^j-k|\right)\phi(0,t)\,dt$$

$$\geq \int_{\mathbb{R}^+} |u_{jn}(x,t_{j+1})-k|\phi(x,t_{j+1}) - |u_{jn}(x,t_j)-k|\phi(x,t_j)\,dx.$$

Then summing on j

$$\int_{\Pi_T^+} |u_n-k|\phi_t + sgn(u_n-k)(f(u_n)-f(k))\phi_x - \frac{f(u_n(0,\cdot))}{u_n(0,\cdot)+c}|u_n-k|\phi_x\,dtdx$$

$$+\int_0^T\left(sgn(u_n(0,\cdot)-k)(f(u_n(0,\cdot))-f(k)) - \frac{f(u_n(0,\cdot))}{u_n(0,\cdot)+c}|u_n(0,\cdot)-k|\right)\phi(0,t)\,dt$$

$$\geq \sum_{j=1}^{k(n)}\int_{\mathbb{R}^+}(|u_{(j-1)n}(x,t_j)-k| - |P_n(u_{(j-1)n})(x,t_j)-k|)\phi(x,t_j)\,dx. \tag{4.10}$$

Consider inequality (4.10) for the converging subsequence given in (i), (ii) and take the limit. Then the first member of the inequality (4.10) tends to

$$\int_{\Pi_T^+} |u-k|\phi_t + sgn(u-k)(f(u)-f(k))\phi_x - \frac{f(w)}{w+c}|u-k|\phi_x\,dtdx \tag{4.11}$$

$$+\int_0^T sgn(w-k)(f(w)-f(k))\phi_x - \frac{f(w)}{w+c}|w-k|\phi\,dt.$$

Moreover (see, for example, [Le2], [Te2]) we see that

$$\limsup_{n\to\infty}\sum_{j=1}^{k(n)}\int_{\mathbb{R}^+}(|u_{(j-1)n}(x,t_j)-k| - |P_n(u_{(j-1)n})(x,t_j)-k|)\phi(x,t_j)\,dx \geq 0.$$

Therefore the expression (4.11) is positive. To conclude the proof, we have to prove that $u(0,t) = w(t)$ for almost every $t \in (0,T)$. Using a suitable test function we obtain

$$sgn(w(t)-k)(f(w(t))-f(k)) - \frac{f(w(t))}{w(t)+c}|w(t)-k|$$

$$\geq sgn(u(0,t)-k)(f(u(0,t))-f(k)) - \frac{f(w(t))}{w(t)+c}|u(0,t)-k| \qquad \text{for almost every } t.$$

Choosing $k > \max\{u(0,t), w(t)\}$ we have

$$f(u(0,t)) - f(w(t)) \geq \frac{f(w(t))}{w(t)+c}(u(0,t)-w(t)),$$

whereas $k < \min\{u(0,t), w(t)\}$ gives

$$f(u(0,t)) - f(w(t)) \le \frac{f(w(t))}{w(t)+c}(u(0,t) - w(t)).$$

Thus we conclude $\frac{f(w(t))}{w(t)+c} = \frac{f(u(0,t))}{u(0,t)+c}$, since w and u are in $\mathcal{S}(c)$ this implies $w(t) = u(0,t)$ for almost any t. □

Observe that the uniqueness result in Proposition 4.1 ensures that all the sequence $\{u_n\}$, constructed in the previous proof, converges to the solution u.

5. Initial data of Class B. In this section we prove Theorem 2.2 for initial data of Class B. Due to Theorem 2.1, any solution of problem (FB) now takes values in the interval $[q, F_c(p)]$.

PROPOSITION 5.1. *Let* (1.3) *hold; let* u_0 *be of Class* B. *Then there exists a unique entropy solution of problem* (FB) *with the free boundary given by* $s(t) = \frac{f(p)}{p+c}t$.

Proof. Since $u(s(t),t) \in \{p,q\} \cap \mathcal{S}(c)$, we have $s'(t) = \frac{f(p)}{p+c}$ $(= \frac{f(q)}{q+c}$ if $q \in \mathcal{S}(c))$.

Let $\overline{u}$ be the entropy solution of the following free boundary problem:

$$\begin{cases} \overline{u}_t + f(\overline{u})_x - \dfrac{f(p)}{p+c}\overline{u}_x = 0 & \text{in} \quad \Pi_T^+, \\ \overline{u}(x,0) = u_0(x) & \text{in} \quad \mathbb{R}^+ \times \{0\}, \\ \overline{u}(0,t) = p & \text{in} \quad \{0\} \times (0,T). \end{cases}$$

Let us prove that the function $u = \overline{u}(x - s(t), t)$ is an entropy solution of the free boundary problem (FB). This follows by proving that for almost every $t \in ((0,T))$ $\overline{u}(0,t) \in \{q, p\}$ when $q \in \mathcal{S}(c)$ and $\overline{u}(0,t) = p$ when $q \notin \mathcal{S}(c)$.

For the compatibility condition along the boundary we have for almost every $t \in (0,T)$

$$\frac{f(\overline{u}(0,t)) - f(k)}{\overline{u}(0,t) - k} \le \frac{f(p)}{p+c} \qquad \text{for every } k \in (\overline{u}(0,t), p].$$

If $q \in \mathcal{S}(c)$ and $\overline{u}(0,t) = q$ the previous inequality holds. Otherwise, reasoning as in the proof of Proposition 4.2 we obtain $\overline{u}(0,t) = p$. When $q \notin \mathcal{S}(c)$ we necessarily have $\overline{u}(0,t) = p$.

We prove uniqueness using the estimate (4.1). In such a case the free boundary is a priori known. Consider two solutions (s, u_1) and (s, u_2). Let u_1^s, u_2^s be the associate functions (see (3.2)) defined in Π_T^+. We obtain

$$\int_{\mathbb{R}^+} |u_1^s(x,t) - u_2^s(x,t)|\, dx$$

$$\le \int_0^t sgn(u_1^s(0,\tau) - {u_2}^s(0,\tau))(f(u_1^s(0,\tau)) - f(u_2^s(0,\tau)) - s'(\tau)(u_1^s(0,\tau) - u_2^s(0,\tau)))\, d\tau$$

$$= 0.$$

Since for almost every $t \in (0,T)$, we have

$$\frac{f(u_1^s(0,t))}{u_1^s(0,t)+c} = \frac{f(u_2^s(0,t))}{u_2^s(0,t)+c} = s'(t). \qquad \square$$

6. Initial data of Class C. In this last section we prove Theorem 2.2 for initial data in Class C, using the results of sections 4 and 5 and the finite speed of propagation of the characteristics.

PROPOSITION 6.1. *Let* (1.3) *hold; let* $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ *be of Class* C. *Then for any entropy solution* u *of the free boundary problem* (FB), $u(\cdot, t)$ *is in the Class* C *for any* t *in* $(0,T)$.

Proof. This can be shown by observing that any entropy solution of the free boundary problem (FB) is the solution of an opportune mixed value problem (Theorem 2.1) that can be thought in Π_T^+ by Proposition B.1. Then using a Godunov-type approximation method for mixed value problems (as in the proof of Proposition 4.4), we can approximate the solution u of the problem in the $C([0,T], L^1_{loc}(\mathbb{R}^+))$ norm by functions having a bounded number of oscillations for any fixed t (solving Riemann problem the number of oscillations do not increase). Therefore the solution u has, for any fixed t, a finite number of oscillations outside a set of null measure. □

LEMMA 6.1. *Let* v *be a function from* $\mathbb{R}^+$ *to* $\mathbb{R}^+$ *belonging to Class* C. *Let* $\{I_k\}$, $\{k = 0, \dots, n\}$, *some intervals of* $\mathbb{R}^+$ *such that* $I_j \neq I_l$ *when* $j \neq l$, *and* $\mathbb{R}^+ = \bigcup_{k=0}^n I_k$. *Then there exists* $k \in \{0, \dots, n\}$ *and* $\epsilon > 0$ *such that* $v((0,\epsilon)) \subset I_k$ *for almost any* $x \in (0,\epsilon)$.

Proof. Consider the interval $[0,1)$. There exists k_1 such that $\mu\{x \in [0,1) : v(x) \in I_{k_1}\} > 0$, where μ is the Lebesgue measure; hence there exists j_1 such that

$$\mu(\{x \in [2^{-j_1}, 1) : v(x) \in I_{k_1}\}) > 0.$$

If

$$\mu\left(\left\{x \in [0, 2^{-j_1}) : v(x) \in \bigcup_{l \neq k_1} I_l\right\}\right) = 0$$

the claim is proved; otherwise there exists $k_2 \neq k_1$ such that

$$\mu(\{x \in [0, 2^{-j_1}) : v(x) \in I_{k_2}\}) > 0,$$

and analogously we can find j_2 such that

$$\mu(\{x \in [2^{-j_2}, 2^{-j_1}) : v(x) \in I_{k_2}\}) > 0.$$

Suppose that the claim is false. Then proceeding as above, we can construct a sequence of integers $\{j_s\}$, $j_s < j_{s+1}$ and a sequence of sets $\{I_{k_s}\}$, $k_s \neq k_{s+1}$ such that

$$\mu(\{x \in [2^{-j_s}, 2^{-j_{s-1}}) : v(x) \in I_{k_s}\}) > 0.$$

Since the interval I_k are only $n+1$, this implies that v does not belong to the Class C, hence the contradiction. □

PROPOSITION 6.2. *Let* $u_0 \in L^\infty(\mathbb{R}^+) \cap BV(\mathbb{R}^+)$ *be an initial data in Class* C. *Then there exists at most one entropy solution of problem* (FB).

Proof. Suppose there exists two entropy solutions (u_1, s_1), (u_2, s_2). Let

$$t_1 = \sup\{t \in [0,T) : s_1 \equiv s_2 \text{ in } (0,t)\}.$$

Let us show that $t_1 = T$. Suppose that $t_1 < T$; we see that $u_1(\cdot, t_1) = u_2(\cdot, t_1)$, and there exists a finite number of points n, $0 = p_0 \leq p_1 < p_2 < \cdots < p_n = +\infty$ such

that $S(c) = \cup_{i=1}^{n}[p_{2i}, p_{2i+1}]$ (Proposition 3.1 (iii) and Proposition 3.2 (iii)). From Proposition 6.1 the function $u_1(\cdot, t_1)$ is in Class C. Therefore, from Lemma 6.1 there exists $i \in \{0, \dots, n-1\}$ and $\epsilon > 0$ such that

$$u_1(\cdot, t_1) \in (p_i, p_{i+1}] \quad \text{a.e. in } (s_1(t_1), s_1(t_1) + \epsilon).$$

In this way, near the boundary we are either in the situation of Proposition 4.1 or in that of Proposition 5.1.

Let us change $u_1(\cdot, t_1)$ outside the interval $(s_1(t_1), s_1(t_1) + \epsilon)$ in such a way that the function takes values in the interval $(p_i, p_{i+1}]$ and preserves its regularity. Let us call this function $\tilde{u}_1(\cdot, t_1)$.

To simplify the notations we assume that $t_1 = 0$.

Let

$$M = \sup_{p \in \mathbb{R}^+} |f'(p)|,$$

where M is an upper bound for the speed of the characteristics. Let

$$C_i := \Lambda(s_i(t), T) \cap \{(x, t) \in \Pi_T^+ \,:\, x - s_i(t) < \epsilon - Mt\}, \qquad i = 1, 2.$$

We see that $u_1 \equiv u_2$ in $C_1 \cap C_2$. In fact, let v be the solution of the following mixed value problem:

$$\begin{cases} v_t + f(v)_x = 0 & \text{in } \Lambda(s_1(t), T), \\ v(x, 0) = \tilde{u}_1(x, 0) & \text{in } \mathbb{R}^+ \times \{0\}, \\ v(s_1(t), t) = u_1(s_1(t), t). \end{cases}$$

From the finite speed of propagation of characteristics $u_1 \equiv v$ in C_1. Then the function v is a solution of the free boundary problem.

On the other hand, let w be the solution of the problem

$$\begin{cases} w_t + f(w)_x = 0 & \text{in } \Lambda(s_2(t), T), \\ w(x, 0) = \tilde{u}_1(x, 0) & \text{in } \mathbb{R}^+ \times \{0\}, \\ w(s_2(t), t) = u_2(s_2(t), t). \end{cases}$$

In the same way we see that $w \equiv u_2$ in C_2. Therefore w is an entropy solution of the free boundary problem. Hence, from Propositions 4.1 and 5.1 we have $(v, s_1) \equiv (w, s_2)$. Then $u_1 \equiv u_2$ in $C_1 \cap C_2$ and there exists a positive δ such that $s_1 \equiv s_2$ in $(0, \delta)$. This is a contradiction with the definition of t_1. □

PROPOSITION 6.3. *Let $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$ be an initial data in Class* C. *Then there exists an entropy solution of the free boundary problem* (FB) *in $\mathbb{R}^+ \times \mathbb{R}^+$.*

Proof. Since u_0 is in Class C, we can construct a local entropy solution of problem (FB). In fact, as in the proof of Proposition 6.2 there exists an $\epsilon > 0$ such that for almost any $x \in (0, \epsilon)$, $u_0(x) \in (p_i, p_{i+1}]$ using the existence theorem of the previous sections and the finite propagation of characteristics, we obtain that there exists the entropy solution of the free boundary problem until a time $\bar{t} > 0$.

Let

$$T := \sup\{t \in \mathbb{R}^+ : \text{ there exists a solution of the free boundary problem in } \Pi_t^+\}.$$

Then $T > 0$. Assume $T < \infty$, and define a solution $\overline{u}$ in Π_T^+ such that $\overline{u}(\cdot, t) = u(\cdot, t)$, where for every $t < T$ function u is a solution of the free boundary problem in $\Pi_{t+\delta}^+$ $(t + \delta < T)$.

This function is well defined for the uniqueness result and is a solution of the free boundary problem (FB). Therefore $\overline{u}$ verifies for time T the condition of the Case C. Then $T = +\infty$ by standard continuation arguments.

Proof of Theorem 2.2. The proof follows by Propositions 4.1–4.4, Proposition 5.1, and Propositions 6.2–6.3.

Appendix A. As we said in section 1, the present study is motivated by a model introduced in [Ro] (see also [Fr]). In this appendix we introduce briefly the model for convenience to the reader.

Consider a homogeneous material exposed to ion beams (this procedure is called *ion etching*). Denote by $y = y(x,t)$ the surface of the material at time t. This function satisfies the following Hamilton–Jacobi equation:

$$y_t = -f(y_x)\,, \tag{A.1}$$

where the function f—referred to as *sputtering function*—depends in general on the material.

It is interesting for the applications to study the situation when different materials with different "sputtering functions" are masked together. For example, in the construction of semiconductor devices the technique of ion etching is used to shape the material in a proper way. To this aim, the semiconductor material is masked with photoresistant materials.

In [Ro] the problem of two materials with surface equation, respectively, y_1, y_2, was considered (see Figure 1). Such materials are separated by an interface which is given by a function $g(x)$. Using the continuity condition along the interface we have

$$y_1(s(t),t) = g(s(t)) = y_2(s(t),t), \tag{A.2}$$

where the function $s(t)$ denotes the separation point of the materials at time t.

Let f^1 (respectively, f^2) be the sputtering functions of the materials with surface equation $y_1(x,t)$ (respectively, $y_2(x,t)$). Let $p_1 = y_{1x}$, $p_2 = y_{2x}$, $s(0) = 0$. Deriving the Hamilton–Jacobi equation (A.1) for each material the problem can be written as follows:

$$\begin{cases} p_{2t} + f^2(p_2)_x = 0 & \text{in } \Lambda(s(\cdot),T), \\ p_{1t} + f^1(p_1)_x = 0 & \text{in } \mathbb{R}\times(0,T)\setminus\Lambda(s(\cdot),T), \\ \dfrac{f^1(p_1(s(t),t))}{p_1(s(t),t) - g'(s(t))} = s'(t) = \dfrac{f^2(p_2(s(t),t))}{p_2(s(t),t) - g'(s(t))}, & \\ p_1(x,0) = p_1^0(x) & \text{in } \mathbb{R}^-\times\{0\}, \\ p_2(x,0) = p_2^0(x) & \text{in } \mathbb{R}^+\times\{0\}, \\ s(0) = 0. & \end{cases} \tag{A.3}$$

Observe that the interface condition is obtained by deriving (A.2).

Problem (A.3) is a free boundary problem in which the unknowns are the functions s, p_1, p_2.

Let us point out that problem (A.3) is a generalization of a Cauchy problem ($g' = \infty$, $s(t) \equiv 0$)

$$\begin{cases} u_t + F(u,x)_x = 0 & \text{in } \mathbb{R}\times\mathbb{R}, \\ u(x,0) = u_0(x), & \end{cases}$$

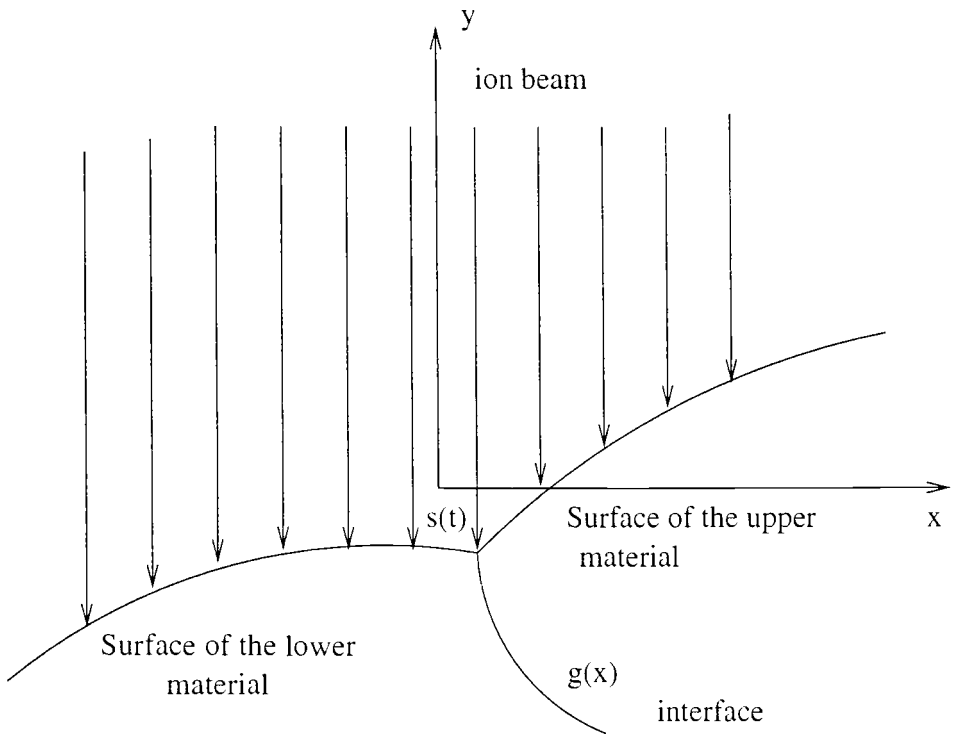


FIG. 1.

where

$$F(u,x) = \begin{cases} f^1(u) & \text{per } x < 0, \\ f^2(u) & \text{per } x \geq 0. \end{cases}$$

This problem arises in several applications: in continuous sedimentation of solid particles in a liquid [GR], [DW], [Ca], in two phase flow in porous media [GHR], and in traffic flow analysis [Mo].

To simplify the free boundary problem (A.3) we suppose that the second material lays over the first material (i.e., $g'(x) < 0$). In this way, following [Ro], by physical considerations, we expect that the evolution of the boundary is affected only by the second material.

Therefore, we can divide problem (A.3) into two separate problems. In the first one, we look for s, p_2 such that

$$\text{(A.4)} \qquad \begin{cases} p_{2t} + f^2(p_2)_x = 0 & \text{in } \Lambda(s(\cdot), T), \\ \dfrac{f^2(p_2(s(t),t))}{p_2(s(t),t) - g'(s(t))} = s'(t), & \\ p_2(x,0) = p_2^0(x) & \text{in } \mathbb{R}^+ \times \{0\}. \end{cases}$$

In the second one we search for p_1 such that

$$\text{(A.5)}\qquad \begin{cases} p_{1t} + f^1(p_1)_x = 0 & \text{in } \mathbb{R}\times(0,T)\setminus\Lambda(s(\cdot),T),\\ \dfrac{f^1(p_1(s(t),t))}{p_1(s(t),t)-g'(s(t))} = s'(t), & \\ p_1(x,0) = p_1^0(x) & \text{in } \mathbb{R}^-\times\{0\}. \end{cases}$$

Observe that problem (A.4) is the free boundary problem (FB), while problem (A.5) is a moving boundary problem. The latter can be easily solved using the existing theory of initial-boundary value problems for scalar conservation laws, when the interface $s(t)$ is known solving the previous problem (see [BLN], [Te1], [Te2]).

Appendix B. In this appendix we collect some results concerning entropy solutions of initial-boundary value problems for conservation laws in noncylindrical domains. We refer the reader to [Te1] for a more complete discussion of this topic.

We denote by $u(s(t),t)$ the trace of the BV function u along the (smooth) curve $s = s(t)$.

DEFINITION B.1. *Let $s \in Lip(0,T)$, $s(0) = 0$. Let $u_0 \in BV_{loc}(\mathbb{R}^+)\cap L^\infty(\mathbb{R}^+)$, $a_0 \in BV((0,T))$. We say that $u \in BV_{loc}(\Lambda(s(\cdot),T))\cap L^\infty(\Lambda(s(\cdot),T))$ is an entropy solution of the problem*

$$\text{(B.1)}\qquad \begin{cases} u_t + f(u)_x = 0 & in \quad \Lambda(s(\cdot),T),\\ u(x,0) = u_0(x) & in \quad \mathbb{R}^+\times\{0\},\\ u(s(t),t) = a_0(t) & in \quad (0,T), \end{cases}$$

if (i) *for any $\phi \in C_0^1(\Lambda(s(\cdot),T))$, $\phi \geq 0$, and any $k \in \mathbb{R}$*

$$\int_{\Lambda(s(\cdot),T)} |u-k|\phi_t + sgn(u-k)(f(u)-f(k))\phi_x\,dxdt + \int_{\mathbb{R}^+} |u_0-k|\phi(x,0)\,dx \geq 0\,;$$

(ii) *for almost every $t \in (0,T)$ the trace $u(s(t),t)$ verifies*

$$\frac{f(u)-f(k)}{u-k} \leq s' \qquad k \in [\min\{a_0(t),u\},\max\{a_0(t),u\}]\setminus\{u\}\,.$$

According to the above definition, the boundary condition in problem (A.1) is not assumed in the strong sense—instead, it is only a *compatibility condition* for the valued assumed by the solution on the boundary. More precisely, condition (ii) in the previous definition means that the admissible discontinuities between the boundary value and the trace of the solution are those for which the characteristics of the discontinuity point outwards.

It is worth noting that we can reduce ourselves to an initial-boundary problem in a cylindric domain by a suitable change of coordinates. We associate with any given function $s \in Lip(0,T)$ and $u \in BV(\Lambda(s(\cdot),T))\cap L^\infty(\Lambda(s(\cdot),T))$, the function

$$\text{(B.2)}\qquad u^s(x,t) = u(x+s(t),t)$$

defined in $\Pi_T^+ = \mathbb{R}^+\times(0,T)$.

PROPOSITION B.1. *The function u is an entropy solution of problem* (B.1) *if and only if the associate function $u^s(y,t) = u(y + s(t), t)$ is an entropy solution of the problem*

$$\text{(B.3)} \qquad \begin{cases} u^s_t + f(u^s)_y - s'(t)u^s_y = 0 & \text{in} \quad \Pi^+_T, \\ u^s(y,0) = u_0(y) & \text{in} \quad \mathbb{R}^+ \times \{0\}, \\ u^s(0,t) = a_0(t) & \text{in} \quad \{0\} \times (0,T). \end{cases}$$

This equivalence follows easily by changing the x coordinate in the integral formulation of the problem (B.1) and by observing that it is not restrictive to use Lipschitz continuous functions vanishing at the boundary instead of C^1_0 functions.

We refer the reader to [BLN] for existence and uniqueness results concerning initial-boundary problems.

Let us mention the following comparison result that is used in the paper.

THEOREM B.1. *Let u_0, $v_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+)$. Let u (respectively, v) be the entropy solution of problem* (A.3) *with initial data u_0, a_0 (respectively, v_0, b_0). Then for any fixed $R > 0$ the following inequality holds for almost any $t \in (0,T)$:*

$$\text{(B.4)} \qquad \int_0^R [v(x,t) - u(x,t)]_+ \, dx \le \int_0^{R+Mt} [v_0(x) - u_0(x)]_+ \, dx$$

$$+ \int_0^t H(v(0,\tau) - u(0,\tau))(f(v(0,\tau)) - f(u(0,\tau)) - s'(\tau)(v(0,\tau) - u(0,\tau))) \, d\tau,$$

where H is the standard Heaviside function and

$$\text{(B.5)} \qquad M := \sup_{|u| \le \max(\|u\|_{L^\infty}, \|v\|_{L^\infty})} |f'(u)| + \|s'\|_{L^\infty(0,T)}.$$

Moreover if $u_0 \in BV(\mathbb{R}^+) \cap L^\infty(\mathbb{R}^+) \cap L^1(\mathbb{R}^+)$, any entropy solution of the problem (B.3) *verifies the conservation equality*

$$\text{(B.6)} \qquad \int_{\mathbb{R}^+} u^s(y,t) \, dx = \int_{\mathbb{R}^+} u_0(y) \, dx + \int_0^t (f(u^s(0,\tau)) - f(0)) \, d\tau - \int_0^t s'(\tau) u(0,\tau) \, d\tau.$$

For a proof of this result see [Te1, Theorem 1.1 and Proposition 3.4], where actually we considered only flux functions $f(u)$. The extension to flux of the type $f(u) + h(t)u$ with $h(t) \in L^\infty(0,T)$ follows by simple modifications.

The following comparison theorem is a consequence of the previous one (see again [Te1]).

THEOREM B.2. *Assume the hypothesis of Theorem* B.1. *Then for almost every $t \in (0,T)$ there holds*

$$\int_0^R [u(x,t) - v(x,t)]_+ dx \le \int_0^{R+Mt} [u_0(x) - v_0(x)]_+ dx$$

$$+ M \int_0^t [a_0(\tau) - b_0(\tau)]_+ d\tau,$$

where the constant M is given by (B.5).

REFERENCES

[BLN] C. Bardos, A.Y. Le Roux, and J.C. Nedelec, *First order quasilinear equations with boundary condition*, Comm. Partial Differential Equations, 4 (1979), pp. 1017–1034.

[Ca] J. Ph. Chancelier, *Analysis of a conservation PDE with discontinuous flux: A model of settler*, SIAM J. Appl. Math., 54 (1994), pp. 954–995.

[Da] I.I. Danilyuk, *On the Stefan problem*, Russian Math. Surveys, 40 (1985), pp. 157–223.

[Df] C. Dafermos, *Polygonal approximation of solutions of the initial value problem for a conservation law*, J. Math. Anal. Appl., 38 (1972), pp. 33–41.

[DW] S. Diehl and N.O Wallin, *Scalar conservation laws with discontinuous flux function* II. *On the stability of the viscous profiles*, Comm. Math. Phys., 176 (1996), pp. 45–71.

[EG] L. Evans and R. Gariepy, *Measure Theory and Fine Properties of Functions*, Stud. Adv. Math. 5, CRC Press, Boca Raton, FL, 1992.

[FH] A. Friedman and B. Hu, *The Stefan problem for a hyperbolic heat equations*, J. Math. Anal. Appl., 37 (1984), pp. 187–279.

[Fr] A. Friedman, *Mathematics in Industrial Problems*, IMA Vol. Mathematics Appl. 16, Springer–Verlag, New York, Berlin, 1988.

[Ge] J. Gerlach, *Two linearized models for a hyperbolic free boundary value problem*, Z. Angew. Math. Phys., 35 (1984), pp. 181–192.

[GHR] T. Gimse, N.H. Risebro, and N. Henrik, *Riemann problems with a discontinuous flux function*, in Third International Conference on Hyperbolic Problems I, II, Studentlitteratur, Lund, Sweden, 1991, pp. 488–502.

[GR] T. Gimse and N.H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.

[Hi] C. Denson Hill, *A hyperbolic free boundary problem*, J. Math. Anal. Appl., 31 (1970), pp. 117–129.

[Kr] S.N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.

[KM] V.M. Kirilich and A.D. Myshkis, *A generalized semilinear hyperbolic Stefan problem on the real line*, Differential Equations, 27 (1991), pp. 356–360.

[La] P. Lax, *Hyperbolic system of conservation laws*, Comm. Pure Appl. Math., 10 (1957), pp. 537–556.

[Le1] A.Y. Le Roux, *Etude du problème mixte pour une équation quasilinéaire du premier ordre*, C.R. Acad. Sci. Paris Sér. I Math., 285 (1977), pp. 351–354.

[Le2] A.Y. Le Roux, *Cours de D.E.A.* 1984–85, Université de Bordeaux I, 1985.

[Li] Li Ta-Tsien, *Global solutions to some free boundary problems for quasilinear hyperbolic systems and applications*, in Nonlinear Hyperbolic Problems (St. Etienne, 1986), Lecture Notes in Math. 1270 , Springer, Berlin, New York, 1987, pp. 195–209.

[LY] Li Ta-Tsien and Yu Wen-Ci, *Boundary value problems for quasilinear hyperbolic systems*, Duke University Mathematics Series V, Durham, NC, 1985.

[Ma1] A. Majda, *The stability of multidimensional shock fronts*, Memoirs of AMS, 275 (1983).

[Ma2] A. Majda, *The existence of multidimensional shock fronts*, Memoirs of AMS, 281 (1983).

[Me] G. Metivier, *Interaction de deux chocs pour un systeme de deux lois conservation, en dimension deux d'espace*, Trans. Amer. Math. Soc., 296 (1986), pp. 431–479.

[Mo] S. Mochon, *An analysis of the traffic on highways with changing surface conditions,* Math. Model., 9 (1987), pp. 1–11.

[Ol] O. Oleinik, *Discontinuous solutions of conservation laws*, AMS Transl., 2 (1964), pp. 95–170.

[Ro] D.S. Ross, *Two new boundary problems for scalar conservation laws*, Comm. Pure Appl. Math., 41 (1988), pp. 725–737.

[SAWD] A.D. Solomon, V. Alexiades, D.G. Wilson, and J. Drake, *On the formulation of hyperbolic Stefan problems*, Quart. Appl. Math., 43 (1985), pp. 295–304.

[SG] L.M. de Socio and G. Gualtieri, *A hyperbolic Stefan problem*, Quart. Appl. Math., 41 (1983), pp. 253–259.

[Sh] N.V. Shemetov, *Existence and stability results for the hyperbolic Stefan problem with relaxation*, Ann. Mat. Pura Appl., 168 (1995), pp. 301–316.

[SW] R.E. Showalter and N.J. Walkington, *A hyperbolic Stefan problem*, Quart. Appl. Math., 45 (1987), pp. 769–781.

[Te1] A. Terracina, *Comparison properties for scalar conservation laws with boundary conditions*, Nonlinear Anal., 28 (1997), pp. 633–653.

[Te2] A. Terracina, *Applicazioni di Teoremi di Confronto per Leggi di Conservazione con Condizioni al Bordo*, Tesi di dottorato, Universitá degli Studi di Roma "La Sapienza," Rome, Italy, 1998.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1010–1028

DETERMINATION OF A THIRD-ORDER OPERATOR FROM TWO OF ITS SPECTRA*

L. AMOUR†

Abstract. We consider a complex third-order differential operator on a bounded interval with boundary conditions presenting a mixed aspect of the Dirichlet and the periodic problems. It is proved that two spectra are sufficient to determine the operator. This result is valid under applying readily verifiable hypotheses simultaneously to the two spectra.

Key words. inverse spectral theory, third-order operator, determination of coefficients

AMS subject classifications. 34A55, 34B05, 34L40, 47E05

PII. S0036141097329639

1. Introduction and result.

Consider the third-order differential operator

$$L_{p,q} = iD^3 + iDq + iqD + p$$

defined on $[0,1]$ with the boundary conditions

$$(DP^+) \quad \begin{cases} y(1) = 0, \\ y'(1) = y'(0), \\ y''(0) = 0. \end{cases}$$

Here $(p,q) \in L^2_{\mathbf{R}}[0,1] \times H^1_{\mathbf{R}}[0,1]$. When $q(0) = 0$ this operator is self-adjoint in $L^2_{\mathbf{C}}[0,1]$ with the scalar product $(f,g) = \int_0^1 f\ \overline{g}\, dx$ and has a discrete spectrum. We denote by $\mu^+(p,q) = \left(\mu_j^+(p,q)\right)_{j\in\mathbf{Z}}$ the increasing sequence of eigenvalues. Each eigenvalue is real and is of multiplicity of at most *two*.

The same considerations hold when $L_{p,q}$ is associated with the boundary conditions

$$(DP^-) \quad \begin{cases} y(1) = 0, \\ y'(1) = -y'(0), \\ y''(0) = 0. \end{cases}$$

Then we denote by $\mu^-(p,q) = \left(\mu_j^-(p,q)\right)_{j\in\mathbf{Z}}$ the increasing sequence of eigenvalues, each eigenvalue being real and of multiplicity of at most *two*.

Let $\lambda^+ = (\lambda_j^+)_{j\in\mathbf{Z}}$ and $\lambda^- = (\lambda_j^-)_{j\in\mathbf{Z}}$ be two sequences of real numbers. We write $\lambda^+ \cap \lambda^- = \phi$ if and only if $\lambda_i^+ \neq \lambda_j^-\ \forall (i,j) \in \mathbf{Z}\times\mathbf{Z}$.

The main result of the paper is as follows.

THEOREM 1.1. *Let $(p,q) \in L^2_R \times H^1_R$ with $q(0) = 0$. Suppose $\mu^+(p,q) \cap \mu^-(p,q) = \phi$. Then*

(i) *all eigenvalues $\mu_j^+(p,q)$ and $\mu_j^-(p,q)$ are of multiplicity one;*

(ii) *the set $(\mu^+(p,q), \mu^-(p,q), q(1))$ determines uniquely (p,q).*

*Received by the editors October 27, 1997; accepted for publication (in revised form) July 23, 1998; published electronically July 22, 1999.

http://www.siam.org/journals/sima/30-5/32963.html

†Laboratoire de Mathématiques, UPRESA 6056, Université de Reims, Moulin de la Housse, B.P. 1039, 51687 Reims Cedex 2, France (laurent.amour@univ-reims.fr).

Part (i) provides a sufficient condition for having a simple spectra for both DP^+ and DP^- boundary conditions: each eigenvalue for the DP^+ boundary conditions is not an eigenvalue for the DP^- boundary conditions. It is proved in Theorem 3.1.

Given μ^+ (resp., μ^-), the additional knowledge of μ^- (resp., μ^+) gives the multiplicity number (one) of all the μ_j^+ (resp., μ_j^-) when $\mu^+ \cap \mu^- = \phi$.

Part (ii) says the following. Let $(p,q) \in L_R^2 \times H_R^1$ and $(\tilde{p},\tilde{q}) \in L_R^2 \times H_R^1$, suppose $\mu^+(p,q) \cap \mu^-(p,q) = \phi$, if $\mu^+(p,q) = \mu^+(\tilde{p},\tilde{q}), \mu^-(p,q) = \mu^-(\tilde{p},\tilde{q}), q(1) = \tilde{q}(1)$, $q(0) = \tilde{q}(0) = 0$; then $p = \tilde{p}$ a.e. in $[0,1]$ and $q = \tilde{q}$ in $[0,1]$.

Theorem 1.1 is not valid, replacing the DP^+ and DP^- boundary conditions with the boundary conditions $y(0) = y(1) = 0,\ y'(1) = y'(0)$, and $y(0) = y(1) = 0,\ y'(1) = -y'(0)$. It easy to check that $L_{p(x),q(x)}$ and $L_{p(1-x),q(1-x)}$ associated with one of these boundary conditions give the same spectrum.

Let us mention here that the indexation of the $\mu_j(p,q)$'s follows from *the counting lemma* (Lemma 3.4). Furthermore we have $\mu^{\pm}(p,q) = \mu^{\pm}(r,s) \Longleftrightarrow \mu_j^{\pm}(p,q) = \mu_j^{\pm}(r,s) \forall j \in \mathbf{Z}$. The counting lemma yields also that only a finite number of eigenvalues are of multiplicity two, with or without the hypothesis $\mu^+(p,q) \cap \mu^-(p,q) = \phi$.

For the second-order case $-y'' + qy = \lambda y$ with boundary conditions of the type $\cos\alpha y(0) + \sin\alpha y'(0) = 0$, $\cos\beta y(1) + \sin\beta y'(1) = 0$, the determination of q from two of its spectra has been extensively studied, giving numerous results through various methods. These are mainly based on Volterra operator transformations and contour integrations, using spectral functions or norming constants (see, for example, [Bo], [Ma], [Ge-Le], [Le-Ga], [Da-Tr]). The problem is solved for first-order differential systems in [Am]. The fourth-order case is also largely studied; see [Ca-Pe-Sc] for recent results and references. Two sequences of spectral data are not sufficient to determine the two coefficients of a self-adjoint fourth-order operator with separated boundary conditions. This can be seen from the exact solution reconstructive method given in [McLa1] and [McLa2]. Let us mention [Gl] and [Ba] where conditions are given for three sequences to be spectral data for a fourth-order self-adjoint boundary value problem with separated boundary conditions.

For the third-order operator the unique result in this direction is given by Leibenzon [Le] in 1966. The equation $y''' + p_1 y + p_2 y = \lambda y$ is associated with three boundary conditions of the type $y(a) + \alpha y'(a) + \beta y''(a) = 0$, $a = 0$ or 1. This is not a self-adjoint problem. It is proved that p_1 and p_2 are determined by two spectra also with the so-called matrix functions. It is supposed as in Theorem 1.1 that the two sets of eigenvalues do not intersect. It is also assumed that some other weight numbers exist. The result is actually generalized for nth-order differential equations, $n > 2$.

In Theorem 1.1 the two boundary conditions are of a different kind than those of Leibenzon. They allow the determination of the operator from only two spectra and use only one of the assumptions of Leibenzon. The proof of Theorem 1.1(ii) is mainly based on the methods of Pöschel and Trubowitz [P-Tr], a contour integration and a counting lemma.

A first generalization of the boundary conditions DP^+ or DP^- should be $y(1) = 0,\ y'(1) = e^{i\phi}y'(0),\ y''(0) = 0$ for some $\phi \in \mathbf{R}/2\pi\mathbf{Z}$ and one may expect that two sequences of spectral data for two different ϕ determine the operator together with the boundary conditions. This is a problem.

The counting lemma gives a rough asymptotic expansion of the $\mu_j^{\pm}(p,q)$'s. Namely, $\mu_j^+(p,q) - (2j\pi)^3$ and $\mu_j^-(p,q) - (2(j+1)\pi)^3$ are bounded as $|j| \to \infty$. More precise asymptotic expansions are given in the following theorem.

THEOREM 1.2. *Suppose* $(p,q) \in L^2_R \times H^1_R$ *with* $q(0)=0$ *and let* $[q]=\int_0^1 q(t)\,dt$*:*

$$\mu_j^+(p,q) = (2j\pi)^3 - 2j\pi[q] + O(1)$$

and

$$\mu_j^-(p,q) = ((2j+1)\pi)^3 - (2j+1)\pi[q] + O(1),$$

as $|j| \to \infty$.

Let us mention the work of McKean [McKe] devoted to the Boussinesq equation but give numerous results in inverse spectral theory for the non-self-adjoint operator $D^3+qD+Dq+p$ for the periodic case and also for the conditions $y(0)=y(1)=y(2)$, p and q being sufficiently small.

This article is organized as follows. In section 2 the estimates used in sections 3 and 4 are established. Then in section 3, Theorem 1.1(i) is proved. Next the counting lemma is derived. This involves the construction of $Z_1(1,\cdot,p,q)$ (resp., $Y_1(1,\cdot,p,q)$), the analytic extensions of the function vanishing at the $\mu_j^+(p,q)$'s (resp., the $\mu_j^-(p,q)$'s). Theorem 1.1(ii) is proved in section 4. Finally, the proof of Theorem 1.2 is given in section 5.

2. Notations, estimates, analyticity. Consider $(p,q) \in L^2_R \times H^1_R$ and let $\lambda \in \mathbf{C}$. The purpose of this section is to give estimates and analyticity results for the fundamental basis of the solutions to $L_{p,q}y = \lambda y$.

We will often use the abbreviated notation $' = \partial/\partial x$.

The functions $y_1(x,\lambda,p,q)$, $y_2(x,\lambda,p,q)$, $y_3(x,\lambda,p,q)$ are defined as the unique solutions to $L_{p,q}y(x) = \lambda y(x)$, for a.e. $x \in [0,1]$, satisfying the initial conditions

$$\begin{pmatrix} y_1(0) & y_2(0) & y_3(0) \\ y_1'(0) & y_2'(0) & y_3'(0) \\ y_1''(0) & y_2''(0) & y_3''(0) \end{pmatrix} = Identity.$$

In particular, every solution to $L_{p,q}y = \lambda y$ can be expressed as $y(x) = a\,y_1(x,\lambda,p,q) + b\,y_2(x,\lambda,p,q) + c\,y_3(x,\lambda,p,q)$, where $(a,b,c) = (y(0),y'(0),y''(0))$.

Since there is no second-order derivative in the definition of $L_{p,q}$, the Wronskian of y_1, y_2, and y_3 is independent of x. Therefore, for a.e. $x \in [0,1]$,

$$\begin{vmatrix} y_1(x) & y_2(x) & y_3(x) \\ y_1'(x) & y_2'(x) & y_3'(x) \\ y_1''(x) & y_2''(x) & y_3''(x) \end{vmatrix} = 1. \tag{2.1}$$

In particular,

$$\forall x_0 \in [0,1], \quad (y(x_0),y'(x_0),y''(x_0)) \neq (0,0,0). \tag{2.2}$$

Let $\omega = e^{\frac{2i\pi}{3}}$ and $k^3 = \lambda$. When $(p,q) \equiv (0,0)$ it is easy to obtain

$$\begin{aligned} y_1(x,\lambda,0,0) &= \frac{1}{3}\left(e^{ikx} + e^{i\omega kx} + e^{i\omega^2 kx}\right), \\ y_2(x,\lambda,0,0) &= \frac{1}{3ik}\left(e^{ikx} + \omega^2 e^{i\omega kx} + \omega e^{i\omega^2 kx}\right), \\ y_3(x,\lambda,0,0) &= \frac{1}{3(ik)^2}\left(e^{ikx} + \omega e^{i\omega kx} + \omega^2 e^{i\omega^2 kx}\right). \end{aligned} \tag{2.3}$$

Note that $\lambda = 0$ is a removable singularity for $y_2(x,\lambda,0,0)$ and $y_3(x,\lambda,0,0)$.

LEMMA 2.1. *We have that*

$$\begin{aligned}
y_1(x,\lambda,0,0) &= \frac{1}{3}\left(4\cos\left(\frac{kx}{2}\right)\cos\left(\frac{\omega kx}{2}\right)\cos\left(\frac{\omega^2 kx}{2}\right) - 1\right.\\
&\quad \left. -4i\sin\left(\frac{kx}{2}\right)\sin\left(\frac{\omega kx}{2}\right)\sin\left(\frac{\omega^2 kx}{2}\right)\right),\\
y_2(x,\lambda,0,0) &= \frac{1}{3ik}\left(4\cos\left(\frac{kx}{2}\right)\cos\left(\frac{\omega kx}{2}-\frac{\pi}{3}\right)\cos\left(\frac{\omega^2 kx}{2}+\frac{\pi}{3}\right) - 1\right.\\
&\quad \left. -4i\sin\left(\frac{kx}{2}\right)\sin\left(\frac{\omega kx}{2}-\frac{\pi}{3}\right)\sin\left(\frac{\omega^2 kx}{2}+\frac{\pi}{3}\right)\right),\\
y_3(x,\lambda,0,0) &= \frac{1}{3(ik)^2}\left(4\cos\left(\frac{kx}{2}\right)\cos\left(\frac{\omega kx}{2}+\frac{\pi}{3}\right)\cos\left(\frac{\omega^2 kx}{2}-\frac{\pi}{3}\right) - 1\right.\\
&\quad \left. -4i\sin\left(\frac{kx}{2}\right)\sin\left(\frac{\omega kx}{2}+\frac{\pi}{3}\right)\sin\left(\frac{\omega^2 kx}{2}-\frac{\pi}{3}\right)\right).
\end{aligned} \tag{2.4}$$

Lemma 2.1 is elementary but fundamental in achieving all estimates in the *whole* complex plane. Combining (2.4) and the inequality $|\sin z| \le e^{|\Im m z|}$, $z \in \mathbf{C}$, the functions y_1, y_2, and y_3 for $(p,q) \equiv (0,0)$ can be estimated as follows.

LEMMA 2.2. *On* $[0,1]\times\mathbf{C}$,

$$\begin{aligned}
|y_1(x,\lambda,0,0)| &\le 3e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x},\\
|y_2(x,\lambda,0,0)| &\le \frac{3}{|k|}e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x},\\
|y_3(x,\lambda,0,0)| &\le \frac{3}{|k|^2}e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x}.
\end{aligned} \tag{2.5}$$

Let us define for $x \in [0,1]$ and $\lambda \in \mathbf{C}$

$$\Xi(x,\lambda) = e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x}. \tag{2.6}$$

We also have Lemma 2.3.

LEMMA 2.3. *Let* $j = 1,2,3$ *and suppose* $n \le j-1$. *On* $[0,1]\times\mathbf{C}$,

$$\left|\frac{\partial^n}{\partial x^n} y_j(x,\lambda,0,0)\right| \le 3\Xi(x,\lambda). \tag{2.7}$$

Proof. Clearly, for $x \in [0,1]$ and $\lambda \in \mathbf{C}$,

$$y_2(x,\lambda,0,0) = \int_0^x y_1(t,\lambda,0,0)\,dt, \quad y_3(x,\lambda,0,0) = \int_0^x y_2(t,\lambda,0,0)\,dt. \tag{2.8}$$

Equalities in (2.8) show successively with Lemma 2.2 that

$$|y_2(x,\lambda,0,0)| \le \Xi(x,\lambda), \quad |y_3(x,\lambda,0,0)| \le \Xi(x,\lambda). \tag{2.9}$$

Besides,

$$y_3'(x,\lambda,0,0) = y_2(x,\lambda,0,0), \quad y_3'(x,\lambda,0,0) = y_2'(x,\lambda,0,0) = y_1'(x,\lambda,0,0). \tag{2.10}$$

Then (2.9) and (2.10) finish the proof. □

The $L^2_R \times H^1_R$ norm is chosen for convenience as

$$\|(p,q)\|^2_{L^2_R \times H^1_R} = 2\|q\|^2_{L^2_R} + \|q'\|^2_{L^2_R} + \|p\|^2_{L^2_R}.$$

THEOREM 2.4. *For $(x,\lambda,p,q) \in [0,1] \times \mathbf{C} \times L^2_R \times H^1_R$ with $q(0) = 0$,*

$$|y_j(x,\lambda,p,q)| \le \frac{3}{|k|^{j-1}} \Xi(x,\lambda) e^{\sqrt{x}\|(p,q)\|_{L^2_R \times H^1_R}}, \quad j = 1,2,3. \tag{2.11}$$

Proof. For $j = 1,2,3$ we have the integral equation

$$\begin{aligned} y_j(x,\lambda,p,q) = y_j(x,\lambda,0,0) + \int_0^x y_3(x-t,\lambda,0,0)\big(-2q(t)y_j'(t,\lambda,p,q) \\ +(-q'(t) + ip(t))y_j(t,\lambda,p,q)\big)\,dt. \end{aligned} \tag{2.12}$$

Integrating (2.12) by parts gives

$$\begin{aligned} y_j(x,\lambda,p,q) = y_j(x,\lambda,0,0) + \int_0^x \Big(-2y_3'(x-t,\lambda,0,0)q(t) \\ +y_3(x-t,\lambda,0,0)(q'(t) + ip(t))\Big) y_j(t,\lambda,p,q)\,dt \\ - \left[2y_3(x-t,\lambda,0,0)q(t)y_j(t,\lambda,p,q)\right]|_{t=0}^{t=x}. \end{aligned} \tag{2.13}$$

The last term in (2.13) is zero since $q(0) = 0$.

Following Picard's iteration we write

$$y_j(x,\lambda,p,q) = y_j(x,\lambda,0,0) + \sum_{n\ge 1} c_n^j(x,\lambda,p,q), \tag{2.14}$$

where

$$\begin{aligned} c_0^j(x,\lambda,p,q) &= y_j(x,\lambda,0,0), \\ c_n^j(x,\lambda,p,q) &= \int_0^x \Big(-2y_3'(x-t,\lambda,0,0)q(t) + y_3(x-t,\lambda,0,0)(q'(t) \\ &\qquad +ip(t))\Big) c_{n-1}^j(t,\lambda,p,q)\,dt. \end{aligned} \tag{2.15}$$

From (2.15) we have for each $n \ge 1$,

$$\begin{aligned} c_n^j(x,\lambda,p,q) \le \int_{t_1 \le t_2 \le \cdots \le t_{n+1} = x} y_j(t_1,\lambda,0,0) \prod_{m=1}^{n} \Big(-2y_3'(t_{m+1}-t_m,\lambda,0,0) \\ q(t_m) + y_3(t_{m+1}-t_m,\lambda,0,0)(q'(t_m) + ip(t_m))\Big)\,dt_1 \cdots dt_n. \end{aligned} \tag{2.16}$$

From Lemma 2.3

$$|y_3(x,\lambda,0,0)| \le 3\Xi(x,\lambda), \quad |y_3'(x,\lambda,0,0)| \le 3\Xi(x,\lambda), \quad (x,\lambda) \in [0,1] \times \mathbf{C}. \tag{2.17}$$

Then using (2.17) and $|y_j(t_1,\lambda,0,0)| \le 3|k|^{1-j}\Xi(x,\lambda)$ (2.16) gives

$$\begin{aligned}|c_n^j(x,\lambda,p,q)| &\le 3\frac{\Xi(x,\lambda)}{n!|k|^{j-1}}\left(\int_0^x 2|q(t)| + |q'(t)+ip(t)|\,dt\right)^n \\ &\le 3\frac{\Xi(x,\lambda)}{n!|k|^{j-1}}\left(\sqrt{x}\|(p,q)\|_{L_R^2\times H_R^1}\right)^n, \quad n\ge 1. \end{aligned} \tag{2.18}$$

Inequalities of Lemma 2.3 and (2.18) together with (2.14) finish the proof. ∎

THEOREM 2.5. *Let $j = 1,2,3$. For $(x,\lambda,p,q) \in [0,1]\times\mathbf{C}\times L_R^2\times H_R^1$ with $q(0)=0$,*

$$|y_j(x,\lambda,p,q) - y_j(x,\lambda,0,0)| \le \frac{3}{|k|^j}\Xi(x,\lambda)e^{\sqrt{x}\|(p,q)\|_{L_R^2\times H_R^1}}. \tag{2.19}$$

Proof. It is a repetition of the proof of Theorem 2.4 except that in (2.16) we replace once in the product, say, for $m=1$, the inequalities (2.17) by the inequalities

$$|y_3(x,\lambda,0,0)| \le \frac{3}{|k|}\Xi(x,\lambda), \quad |y_3'(x,\lambda,0,0)| \le \frac{3}{|k|}\Xi(x,\lambda), \quad (x,\lambda)\in[0,1]\times\mathbf{C}. \tag{2.20}$$

Inequalities (2.20) can be easily derived in a similar way as those in Lemma 2.3. ∎

THEOREM 2.6. *Let $j=1,2,3$. For each $x\in[0,1]$ and each $(p,q)\in L_R^2\times H_R^1$ with $q(0)=0$,*

$$y_j(x,\lambda,p,q), \quad y_j'(x,\lambda,p,q), \quad y_j''(x,\lambda,p,q) \tag{2.21}$$

are entire functions of λ.

Remark. It can be proved that these functions are also analytic functions of (λ,p,q).

Proof. Fix $j=1,2,3$. The coefficient $c_n^j(x,\lambda,p,q)$, nth term in the power series of $y_j(x,\lambda,p,q)$ is an entire function of λ. The convergence in (2.14) is locally uniform in λ by (2.18). Then $y_j(x,\lambda,p,q)$ is an entire function of λ according to Weierstrass's theorem.

The first derivative with respect to x of the integral equation (2.13) is

$$\begin{aligned}y_j'(x,\lambda,p,q) = y_j'(x,\lambda,0,0) + \int_0^x \Big(&-2y_3''(x-t,\lambda,0,0)q(t) \\ &+ y_3'(x-t,\lambda,0,0)(q'(t)+ip(t))\Big)y_j(t,\lambda,p,q)\,dt. \end{aligned} \tag{2.22}$$

The analyticity of $y_j(x,\lambda,p,q)$ gives the analyticity of $y_j'(x,\lambda,p,q)$. Similarly,

$$\begin{aligned}y_j''(x,\lambda,p,q) = y_j''(x,\lambda,0,0) + \int_0^x \Big(&-2y_3'''(x-t,\lambda,0,0)q(t) \\ &+ y_3''(x-t,\lambda,0,0)(q'(t)+ip(t))\Big)y_j(t,\lambda,p,q)\,dt - 2q(x) \end{aligned} \tag{2.23}$$

proves the analyticity result of $y_j''(x,\lambda,p,q)$.

The estimates for the derivatives of $y_j(x,\lambda,p,q)$ are given in Theorem 2.7.

THEOREM 2.7. *Let $j = 1, 2, 3$. For $(x, \lambda, p, q) \in [0, 1] \times \mathbf{C} \times L^2_R \times H^1_R$ with $q(0) = 0$,*

$$|y'_j(x, \lambda, p, q) - y'_j(x, \lambda, 0, 0)| \leq 3 \frac{\|(p, q)\|_{L^2_R \times H^1_R}}{|k|^{j-1}} \Xi(x, \lambda) e^{\sqrt{x}\|(p,q)\|_{L^2_R \times H^1_R}} \tag{2.24}$$

and

$$|y''_j(x, \lambda, p, q) + 2q(x) - y''_j(x, \lambda, 0, 0)| \leq 3 \frac{\|(p, q)\|_{L^2_R \times H^1_R}}{|k|^{j-2}} \Xi(x, \lambda) \times e^{\sqrt{x}\|(p,q)\|_{L^2_R \times H^1_R}}. \tag{2.25}$$

The proof of (2.24) and (2.25) is a consequence of (2.22) and (2.23), respectively. □

3. Proof of Theorem 1.1(i) and the counting lemma. We begin this section by proving Theorem 1.1(i). It is Theorem 3.1(iii).

THEOREM 3.1. *Let $(p, q) \in L^2_R \times H^1_R$ satisfying $q(0) = 0$.*

(i) *$\mu^+(p, q) \cap \mu^-(p, q) = \phi$ if and only if $y_1(1, \lambda, p, q)$ has no zero in $\mathbf{C}$.*

(ii) *Any $\lambda \in \mathbf{R}$ is a simple eigenvalue of $L_{p,q}$ associated with the DP^+ boundary conditions if and only if*

$$\begin{cases} \Im\mathrm{m}\, y_1(1, \lambda, p, q) = 0 \\ \Re\mathrm{e}\, y_1(1, \lambda, p, q) \neq 0 \end{cases} \quad \text{or} \quad \begin{cases} \Im\mathrm{m}\, y_1(1, \lambda, p, q) = 0 \\ \Re\mathrm{e}\, y'_1(1, \lambda, p, q) \neq 0 \end{cases}.$$

Any $\lambda \in \mathbf{R}$ is a simple eigenvalue of $L_{p,q}$ associated with the DP^- boundary conditions if and only if

$$\begin{cases} \Re\mathrm{e}\, y_1(1, \lambda, p, q) = 0 \\ \Im\mathrm{m}\, y_1(1, \lambda, p, q) \neq 0 \end{cases} \quad \text{or} \quad \begin{cases} \Re\mathrm{e}\, y_1(1, \lambda, p, q) = 0 \\ \Im\mathrm{m}\, y'_1(1, \lambda, p, q) \neq 0 \end{cases}.$$

(iii) *If $\mu^+(p, q) \cap \mu^-(p, q) = \phi$, then $\mu^+(p, q)$ and $\mu^-(p, q)$ are both simple spectra.*

The proof uses the following preliminary observations.

Let $[y, z] = y'z - yz'$.

LEMMA 3.2. *Let $\lambda \in \mathbf{C}$ and suppose $(p, q) \in L^2_R \times H^1_R$. Then*

$$\begin{aligned} [y_3(x, \lambda, p, q), y_2(x, \lambda, p, q)] &= \bar{y}_3(x, \bar{\lambda}, p, q), \\ [y_3(x, \lambda, p, q), y_1(x, \lambda, p, q)] &= \bar{y}_2(x, \bar{\lambda}, p, q), \\ [y_2(x, \lambda, p, q), y_1(x, \lambda, p, q)] &= \bar{y}_1(x, \bar{\lambda}, p, q) - 2q(0)\bar{y}_3(x, \bar{\lambda}, p, q). \end{aligned} \tag{3.1}$$

This result is given by McKean [McKe].

Proof. Check that the left- and right-hand sides of each identity in (3.1) satisfy $L_{-p,q}y = -\lambda y$ and the same initial conditions at $x = 0$. □

Let

$$M^{\pm}(\lambda, p, q) = \begin{pmatrix} y_1(1, \lambda, p, q) & y_2(1, \lambda, p, q) \\ y'_1(1, \lambda, p, q) & y'_2(1, \lambda, p, q) \mp 1 \end{pmatrix}. \tag{3.2}$$

LEMMA 3.3. *Fix* (p,q) *in* $L^2_R \times H^1_R$ *and suppose* $q(0)=0$. *For all* $\lambda \in \mathbf{R}$ *we have*

$$\text{(i)} \quad \begin{cases} y_1(1,\lambda,p,q)=0, \\ \Re\mathrm{e}\, y_1''(1,\lambda,p,q)=0 \end{cases} \iff M^+(\lambda,p,q)=0, \tag{3.3}$$

$$\text{(ii)} \quad \begin{cases} y_1(1,\lambda,p,q)=0, \\ \Im\mathrm{a}\, y_1''(1,\lambda,p,q)=0 \end{cases} \iff M^-(\lambda,p,q)=0. \tag{3.4}$$

Proof. Suppose $q(0)=0$ and fix $\lambda \in \mathbf{R}$. We omit (p,q) from the notation for brevity. From $\bar{y}_1(1,\lambda)=y_2'(1,\lambda)y_1(1,\lambda)-y_2(1,\lambda)y_1'(1,\lambda)$ we have $y_1(1,\lambda)=0 \Rightarrow y_1'(1,\lambda)=0$ or $y_2(1,\lambda)=0$. The equality $\bar{y}_2(1,\lambda)=y_3'(1,\lambda)y_1(1,\lambda)-y_3(1,\lambda)y_1'(1,\lambda)$ gives $y_1(1,\lambda)=y_1'(1,\lambda)=0 \Rightarrow y_2(1,\lambda)=0$. Therefore,

$$y_1(1,\lambda)=0 \Rightarrow y_2(1,\lambda)=0. \tag{3.5}$$

Similarly, $\bar{y}_1'(1,\lambda)=y_2''(1,\lambda)y_1(1,\lambda)-y_2(1,\lambda)y_1''(1,\lambda)$ yields $y_1(1,\lambda)=y_2(1,\lambda)=0 \Rightarrow y_1'(1,\lambda)=0$. Thus, with (3.5) we obtain

$$y_1(1,\lambda)=0 \Leftrightarrow y_1(1,\lambda)=y_1'(1,\lambda)=y_2(1,\lambda)=0.$$

Since $y_1''(1,\lambda)\neq 0$ when $y_1(1,\lambda)=y_1'(1,\lambda)=0$, the proof is finished using $\bar{y}_1''(1,\lambda)=y_2''(1,\lambda)y_1'(1,\lambda)-y_2'(1,\lambda)y_1''(1,\lambda)$. □

Proof of Theorem 3.1. Looking for a solution $y(x)$ to $L_{p,q}y=\lambda y$, for a.e. $x\in[0,1]$ and satisfying the $DP^{\pm}$ boundary conditions, we write

$$y(x)=a\,y_1(x,\lambda,p,q)+b\,y_2(x,\lambda,p,q)+c\,y_3(x,\lambda,p,q)$$

with $(a,b,c)\in \mathbf{C}^3\backslash 0$.

Clearly

$$\begin{cases} a\,y_1(1,\lambda,p,q)+b\,y_2(1,\lambda,p,q)+c\,y_3(1,\lambda,p,q)=0, \\ a\,y_1'(1,\lambda,p,q)+b\,y_2'(1,\lambda,p,q)+c\,y_3'(1,\lambda,p,q)=\pm b, \\ c=0, \end{cases} \tag{3.6}$$

that is to say,

$$M^{\pm}(\lambda,p,q)\begin{pmatrix} a \\ b \end{pmatrix}=\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad (a,b)\neq(0,0).$$

Then any real λ is an eigenvalue of $L_{p,q}$ associated with the $DP^{\pm}$ boundary conditions if and only if $\dim Ker\, M^{\pm}(\lambda,p,q)\geq 1$, i.e., if and only if

$$\det M^{\pm}(\lambda,p,q)=0. \tag{3.7}$$

Using Lemma 3.2 we have $\det M^{\pm}(\lambda,p,q)=\bar{y}_1(1,\lambda,p,q)\mp y_1(1,\lambda,p,q)$. Therefore,

$$\Im m y_1(1,\lambda,p,q)=0 \text{ (resp., } \Re e y_1(1,\lambda,p,q)=0) \tag{3.8}$$

gives the eigenvalues for the DP^+ boundary conditions (resp., for the DP^- boundary conditions). This proves (i).

Any real λ is a *simple* eigenvalue $L_{p,q}$ associated with the $DP^{\pm}$ boundary conditions if and only if dim $Ker\, M^{\pm}(\lambda,p,q)=1$, i.e., if and only if

$$\det M^{\pm}(\lambda,p,q)=0 \quad \text{and} \quad M^{\pm}(\lambda,p,q)\neq \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \tag{3.9}$$

This proves (ii) with Lemma 3.3. Then (iii) follows from (i) and (ii). □

When λ is a *simple* eigenvalue, an immediate consequence of (3.2) and (3.9) is that either the two expressions

$$\begin{aligned} &y_2(1,\lambda,p,q)y_1(x,\lambda,p,q)-y_1(1,\lambda,p,q)y_2(x,\lambda,p,q)\\ &\text{and}\\ &(y_2'(1,\lambda,p,q)\mp 1)y_1(x,\lambda,p,q)-y_1'(1,\lambda,p,q)y_2(x,\lambda,p,q) \end{aligned} \tag{3.10}$$

are equivalent up to a multiplicative constant $\neq 0$ or one (and only one) of the two expressions is identically zero. This provides the eigenfunction (up to a multiplicative constant $\neq 0$) corresponding to λ.

When λ is a *double* eigenvalue, $y_1(\cdot,\lambda,p,q)$ and $y_2(\cdot,\lambda,p,q)$ are two linearly independent eigenfunctions corresponding to λ.

Now we construct explicitly the analytic extensions in $\mathbf{C}$ of $\lambda\mapsto \Im m y_1(1,\lambda,p,q)$ and $\lambda\mapsto \Re e y_1(1,\lambda,p,q)$.

For this, just define the operator

$$l_{p,q}=\begin{pmatrix} p & -D^3-Dq-qD \\ D^3+Dq+qD & p \end{pmatrix}$$

on the *real*-valued function space $\mathrm{L}^2_{\mathbf{R}}[0,1]\times \mathrm{L}^2_{\mathbf{R}}[0,1]$ and consider the eigenvalue problem

$$l_{p,q}\begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix}=\lambda\begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix} \quad \text{for a.e. } x\in[0,1]$$

with either the dl^+ or the dl^- boundary conditions defined by

$$(dl^{\pm}) \quad \begin{cases} Y(1)=Z(1)=0,\\ Y'(1)=\pm Y'(0), \quad Z'(1)=\pm Z'(0),\\ Y''(0)=Z''(0)=0. \end{cases}$$

The operator $l_{p,q}$ with the dl^+ or dl^- boundary conditions is self-adjoint on $\mathrm{L}^2_{\mathbf{R}}[0,1]\times \mathrm{L}^2_{\mathbf{R}}[0,1]$ with the inner scalar product

$$\left\langle \begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix}, \begin{pmatrix} \tilde{Y}(x) \\ \tilde{Z}(x) \end{pmatrix} \right\rangle = \int_0^1 Y(x)\tilde{Y}(x)+Z(x)\tilde{Z}(x)\,dx.$$

Write $y=Y+iZ$, Y and Z being real-valued functions, and observe that $L_{p,q}y=\lambda y$ is equivalent to

$$l_{p,q}\begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix}=\lambda\begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix}$$

for *real* λ. It is different for all complex λ because of right-hand sides.

The spectrum of $L_{p,q}$ with the DL^+ (resp., DL^-) boundary conditions is then the spectrum of $l_{p,q}$ with the dl^+ (resp., dl^-) boundary conditions.

For each $\lambda \in \mathbf{C}$, $\binom{Y_1(x)}{Z_1(x)}$ denotes the solution to

$$l_{p,q}\begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix} = \lambda \begin{pmatrix} Y(x) \\ Z(x) \end{pmatrix}$$

satisfying the initial condition

$$(Y_1 + iZ_1)(0,\lambda,p,q) = 1, \quad (Y_1' + iZ_1')(0,\lambda,p,q) = 0, \quad (Y_1'' + iZ_1'')(0,\lambda,p,q) = 0.$$

We have

$$y_1(x,\lambda,p,q) = Y_1(x,\lambda,p,q) + iZ_1(x,\lambda,p,q) \quad \text{for a.e } x \in [0,1]\ \forall \lambda \in \mathbf{R}. \tag{3.11}$$

Again (3.11) is not satisfied for all $\lambda \in \mathbf{C}$.

Moreover, it is clear that

$$\begin{aligned} \left\{\mu_j^+(p,q), \quad j \in \mathbf{Z}\right\} &= \left\{\lambda \in \mathbf{C}, Z_1(1,\lambda,p,q) = 0\right\}, \\ \left\{\mu_j^-(p,q), \quad j \in \mathbf{Z}\right\} &= \left\{\lambda \in \mathbf{C}, Y_1(1,\lambda,p,q) = 0\right\}. \end{aligned} \tag{3.12}$$

The same considerations (the integral equation and the Picard iteration) as those for $y_1(x,\lambda,p,q)$ in section 2 show that $Y_1(1,\lambda,p,q)$ and $Z_1(1,\lambda,p,q)$ are entire functions of λ (see [Am-Gu, appendix] for a very similar treatment with a first-order system). They are the analytic extensions in $\mathbf{C}$ of $\lambda \mapsto \Im m y_1(1,\lambda,p,q)$ and $\lambda \mapsto \Re e y_1(1,\lambda,p,q)$. Moreover, under the same assumptions of Theorem 2.5, it can be seen that

$$\begin{aligned} &|Y_1(x,\lambda,p,q) - Y_1(x,\lambda,0,0)| \le \frac{3}{|k|}\Xi(x,\lambda)e^{\sqrt{x}\|(p,q)\|_{L^2_R \times H^1_R}} \\ &\text{and} \\ &|Z_1(x,\lambda,p,q) - Z_1(x,\lambda,0,0)| \le \frac{3}{|k|}\Xi(x,\lambda)e^{\sqrt{x}\|(p,q)\|_{L^2_R \times H^1_R}} \end{aligned} \tag{3.13}$$

$\forall \lambda \in \mathbf{C}$.

The final part of this section is concerned with the proof of the counting lemma.

Let

$$\begin{aligned} \Omega_j^- &= \left\{z \in \mathbf{C},\ |\omega^{j-1}z - (2j+1)\pi| \ge \frac{\pi}{2}\right\}, \\ \Omega_j^+ &= \left\{z \in \mathbf{C},\ |\omega^{j-1}z - 2j\pi| \ge \frac{\pi}{2}\right\} \end{aligned} \tag{3.14}$$

for $j = 1,2,3$ and define

$$\Omega^{\pm} = \bigcap_{j=1}^{3} \Omega_j^{\pm}. \tag{3.15}$$

Using this notation we have Lemma 3.4.

LEMMA 3.4 (the counting lemma). *Suppose* $(p,q) \in L^2_{\mathbf{C}} \times H^1_{\mathbf{C}}$ *with* $q(0) = 0$.

(i) *Let the integer* J *satisfy* $(2J+1)\pi > 48e^{\|(p,q)\|}$. *Then* $Z_1(1,\lambda,p,q)$ *has exactly* $2J+1$ *roots, counted with multiplicity in the open k-disc*

$$\{k \in \mathbf{C},\ |k| < (2J+1)\pi\}$$

and exactly one simple root noted $\mu_j^+(p,q)$ *in each open k-disc* $\{k \in \mathbf{C},\ |k-2j\pi| < \pi\}$ *for* $|j| > J$.

(ii) *Let the integer* J *satisfy* $2J\pi > \min(\ln 32, 96e^{\|(p,q)\|})$. *Then* $Y_1(1,\lambda,p,q)$ *has exactly* $2J$ *roots, counted with multiplicity in the open k-disc*

$$\{k \in \mathbf{C},\ |k| < 2J\pi\}$$

and exactly one simple root noted $\mu_j^-(p,q)$ *in each open k-disc* $\{k \in \mathbf{C},\ |k-(2j+1)\pi| < \pi\}$ *for* $|j| \geq J$.

The proof of Lemma 3.4 involves the following inequalities.

LEMMA 3.5.

$$\begin{aligned} &\text{(i) } 96|Y_1(1,\lambda,0,0)| \geq \Xi(1,\lambda) \quad \forall k \in \Omega^- \setminus \{|k| \leq \ln 32\}, \\ &\text{(ii) } 48|Z_1(1,\lambda,0,0)| \geq \Xi(1,\lambda) \quad \forall k \in \Omega^+. \end{aligned} \tag{3.16}$$

The definition of Ξ is given in (2.6).

Proof. (i) Similarly as in Lemma 2.1 it can be seen that

$$Y_1(1,\lambda,0,0) = \frac{1}{3}\left(4\cos\left(\frac{k}{2}\right)\cos\left(\frac{\omega k}{2}\right)\cos\left(\frac{\omega^2 k}{2}\right) - 1\right) \ \forall \lambda \in \mathbf{C}.$$

Then

$$|Y_1(1,\lambda,0,0)|^2 = 16|\Lambda|^2 - 8\Re e\,\Lambda + 1,$$

where

$$\Lambda = \cos\left(\frac{k}{2}\right)\cos\left(\frac{\omega k}{2}\right)\cos\left(\frac{\omega^2 k}{2}\right).$$

Therefore, for any $0 < \varepsilon < 4$,

$$9|Y_1(1,\lambda,0,0)|^2 \geq \varepsilon^2|\Lambda|^2 \tag{3.17}$$

if

$$(16-\varepsilon^2)|\Lambda|^2 - 8|\Lambda| + 1 \geq 0,$$

that is to say, if

$$|\Lambda| \geq (4-\varepsilon)^{-1}. \tag{3.18}$$

Inequality $e^{|\Im m z|} \leq 4|\sin z|$ in $\{z \in \mathbf{C},\ |z| \geq \pi/4\}$ [P-Tr, Lemma 2.1] yields

$$\begin{aligned} &\left|\cos\left(\frac{k}{2}\right)\right| \geq \frac{1}{4}e^{|\Im m\frac{k}{2}|} \geq \frac{1}{4} \ \text{ for } \ k \in \Omega_1^-, \\ &\left|\cos\left(\frac{\omega k}{2}\right)\right| \geq \frac{1}{4}e^{|\Im m\omega\frac{k}{2}|} \geq \frac{1}{4} \ \text{ for } \ k \in \Omega_2^-, \\ &\left|\cos\left(\frac{\omega^2 k}{2}\right)\right| \geq \frac{1}{4}e^{|\Im m\omega^2\frac{k}{2}|} \geq \frac{1}{4} \ \text{ for } \ k \in \Omega_3^-. \end{aligned} \tag{3.19}$$

From (3.19) we have

$$|\Lambda| \geq \frac{1}{4^3}e^{|\Im m\omega\frac{k}{2}|} \quad \text{and} \quad |\Lambda| \geq \frac{1}{4^3}e^{|\Im m\omega^2\frac{k}{2}|} \quad \forall k \in \Omega^-.$$

Consequently

$$|\Lambda| \geq \frac{1}{2}$$

$$\text{if} \quad k \in \left(\left\{ \left| \Im m\omega \frac{k}{2} \right| \geq \ln 32 \right\} \cup \left\{ \left| \Im m\omega^2 \frac{k}{2} \right| \geq \ln 32 \right\} \right) \cap \Omega^- . \tag{3.20}$$

Therefore, (3.18) is satisfied for $\varepsilon = 2$ and for $k \in \Omega^- \setminus \{|k| \leq \ln 32\}$. Then

$$|Y_1(1,\lambda,0,0)| \geq \frac{2}{3}|\Lambda| \quad \forall k \in \Omega^- \setminus \{|k| \leq \ln 32\}, \tag{3.21}$$

and again using (3.19) we have

$$|\Lambda| \geq \frac{1}{4^3}\Xi(1,\lambda) \quad \text{for} \quad k \in \Omega^- . \tag{3.22}$$

Inequalities (3.21) and (3.22) prove (i).

The proof of (ii) is immediate since

$$\begin{aligned} |Z_1(1,\lambda,0,0)| &= \frac{1}{3}\left| 4\sin\left(\frac{k}{2}\right)\sin\left(\frac{\omega k}{2}\right)\sin\left(\frac{\omega^2 k}{2}\right)\right| \quad \forall k \in \mathbf{C} \\ &\geq \frac{1}{48} e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x} \quad \forall k \in \Omega^+ . \end{aligned}$$

□

Proof of Lemma 3.4. (i) Combining (3.13) and (3.16) we obtain

$$|Z_1(1,\lambda,p,q) - Z_1(1,\lambda,0,0)| < |Z_1(1,\lambda,0,0)|$$

$\forall\, k \in \Omega^+$ satisfying $|k| \geq 48e^{\|(p,q)\|}$.

Consequently, Rouché's theorem can be applied on the k-circles $|k| = (2J+1)$ and $|k-2j\pi| = \pi$ for $|j| > J$. Then $Z_1(1,\lambda,p,q)$ and $Z_1(1,\lambda,0,0) = \frac{4}{3}\sin(\frac{kx}{2})\sin(\frac{\omega kx}{2})\sin(\frac{\omega^2 kx}{2})$ have the same number of zeros, counted with multiplicity, in the corresponding open k-discs, the zeros of $Z_1(1,\lambda,0,0)$ being the $(2j\pi)^3$'s, $j \in \mathbf{Z}$.

(ii) Similarly, (3.13) and (3.16) give

$$|Y_1(1,\lambda,p,q) - Y_1(1,\lambda,0,0)| < |Y_1(1,\lambda,0,0)|$$

$\forall\, k \in \Omega^-$ satisfying $|k| \geq \max(\ln 32, 96e^{\|(p,q)\|})$. Then $Y_1(1,\lambda,p,q)$ and $Y_1(1,\lambda,0,0)$ have the same number of zeros in the open k-discs defined in Lemma 3.3(ii), the zeros of $Y_1(1,\lambda,0,0)$ being the $((2j+1)\pi)^3$'s, $j \in \mathbf{Z}$, up to an irrelevant additive $O(e^{-c|k|})$ term, c being a positive numerical constant. □

4. Proof of Theorem 1.1(ii). Let us mention first that the indexing of the $\mu_j^{\mp}(p,q)$'s following Lemma 3.4 has the property

$$\mu^{\pm}(p_1,q_1) = \mu^{\pm}(p_2,q_2) \Longleftrightarrow \mu_j^{\pm}(p_1,q_1) = \mu_j^{\pm}(p_2,q_2) \quad \forall\, j \in \mathbf{Z}.$$

Define

$$y(x,\lambda,p,q) = y_2(1,\lambda,p,q)y_1(x,\lambda,p,q) - y_1(1,\lambda,p,q)y_2(x,\lambda,p,q) \tag{4.1}$$

$\forall \lambda \in \mathbf{C}$ and a.e. $x \in [0,1]$.

Let us recall that under the assumptions of Theorem 1.1, $y(\cdot, \mu_j^{\pm}, p, q)$ is the eigenfunction corresponding to the eigenvalue $\mu_j^{\pm}(p,q)$.

We will often use the abbreviated notation $\dot{\cdot} = \partial/\partial\lambda$. We now have Lemma 4.1.

LEMMA 4.1. *Let* $(p,q) \in L^2_R \times H^1_R$. *Suppose* $q(0) = 0$. *Then*

$$\begin{aligned} \int_0^1 |y(x, \mu_j^+(p,q), p, q)|^2 \, dx &= -2(Y_1 \dot{Z}_1)(1, \mu_j^+(p,q), p, q), \\ \int_0^1 |y(x, \mu_j^-(p,q), p, q)|^2 \, dx &= 2(Z_1 \dot{Y}_1)(1, \mu_j^-(p,q), p, q) \end{aligned} \tag{4.2}$$

$\forall j \in \mathbf{Z}$.

Remark. When $\mu^+(p,q) \cap \mu^-(p,q) = \phi$, the geometric multiplicity of each $\mu_j^+(p,q)$ (resp., $\mu_j^-(p,q)$) is one. Besides, from Lemma 4.1, its algebraic multiplicity, defined as its order as a root in $Z_1(1, \lambda, p, q)$ (resp., $Y_1(1, \lambda, p, q)$), is also one.

Proof. Let $\lambda \in \mathbf{C}$. We have

$$i\dot{y}''' + 2iq\dot{y}' + (iq' + p)\dot{y} = \lambda \dot{y} + y \tag{4.3}$$

and

$$i\bar{\dot{y}}''' + 2iq\bar{\dot{y}}' + (iq' - p)\bar{\dot{y}} = -\bar{\lambda}\bar{\dot{y}}. \tag{4.4}$$

Combining (4.3) and (4.4) we have for $\lambda \in \mathbf{R}$

$$\begin{aligned} |y|^2 \quad &= 2i(q\dot{y}\bar{\dot{y}})' + i\dot{y}'''\bar{\dot{y}} + i\dot{y}\bar{\dot{y}}''' \\ &= i(\dot{y}\bar{\dot{y}}'' - \dot{y}'\bar{\dot{y}}' + \dot{y}''\bar{\dot{y}} + 2q\dot{y}\bar{\dot{y}})'. \end{aligned} \tag{4.5}$$

Suppose $\lambda = \mu_j^{\pm}$. From (4.5), since $q(0) = 0$ and $y(0) = 0$, we obtain

$$\int_0^1 |y|^2 \, dx = i[\dot{y}\bar{\dot{y}}'' - \dot{y}'\bar{\dot{y}}' + \dot{y}''\bar{\dot{y}}]|_{x=0}^{x=1}. \tag{4.6}$$

Besides, it is easy to compute

$$\begin{aligned} &\dot{y}(1) = 0, \quad \dot{y}''(0) = 0, \quad \dot{y}'(0) = -\dot{y}_1(1), \\ &\bar{\dot{y}}'(0) = -\bar{\dot{y}}_1(1), \quad \bar{\dot{y}}'(1) = -y_1(1), \quad \dot{y}'(1) = -\dot{\bar{y}}_1(1), \end{aligned} \tag{4.7}$$

where all functions in (4.7) are evaluated at $\mu_j^{\pm}(p,q)$, $j \in \mathbf{Z}$.

Finally, since

$$\begin{aligned} y_1(1, \lambda, p, q) &= \bar{y}_1(1, \lambda, p, q) \quad \text{for } \lambda = \mu_j^+(p,q), \\ y_1(1, \lambda, p, q) &= -\bar{y}_1(1, \lambda, p, q) \quad \text{for } \lambda = \mu_j^-(p,q), \end{aligned} \tag{4.8}$$

we have from (4.5), (4.6), and (4.7),

$$\int_0^1 |y(x, \lambda, p, q)|^2 \, dx = -iy_1(1, \lambda, p, q) \frac{\partial}{\partial \lambda} \left(y_1(1, \lambda, p, q) \pm \bar{y}_1(1, \lambda, p, q) \right)$$

at $\lambda = \mu_j^{\pm}(p,q)$, $j \in \mathbf{Z}$.

The proof is finished since $\Re e y_1(1, \cdot, p, q)$ coincides with $Y_1(1, \cdot, p, q)$ and $\Im m y_1(1, \cdot, p, q)$ coincides with $Z_1(1, \cdot, p, q)$ on the real axis. □

For $g \in L^2_{\mathbf{R}}[0,1]$, g^* is defined by $g^*(x) = g(1-x)$ for a.e. $x \in [0,1]$.

LEMMA 4.2. *Let* $(p,q) \in L^2_R \times H^1_R$, $\lambda \in \mathbf{C}$ *and suppose* $q(0) = 0$. *Then*

$$\begin{pmatrix} y_1 & y_2 & y_3 \\ y_1' & y_2' & y_3' \\ y_1'' & y_2'' & y_3'' \end{pmatrix}(1,\lambda,p^*,q^*) = \begin{pmatrix} y_3'' & y_3' & y_3 \\ y_2'' & y_2' & y_2 \\ y_1'' & y_1' & y_1 \end{pmatrix}(1,\lambda,p,q). \tag{4.9}$$

Remark. The equality $y_3(1,\lambda,p,q) = y_3(1,\lambda,p^*,q^*)$ may be explained. The eigenvalues corresponding to the boundary conditions $y(0) = y(1) = 0$, $y'(0) = e^{i\phi}y'(1)$ with $\phi \in [0,2\pi)$ are the zeros of $\cos(\phi/2)\Re e y_3(1,\lambda,p,q) - \sin(\phi/2)\Im m y_3(1,\lambda,p,q)$. Take $\phi = 0$. One should expect, using a Weierstrass product for the analytic extension of $\Re e y_3(1,\lambda,p,q)$, that these eigenvalues determine $\Re e y_3(1,\lambda,p,q)$. The sequence of eigenvalues corresponding to $\phi = \pi$ should determine $\Im m y_3(1,\lambda,p,q)$. These two sequences of eigenvalues do not change replacing (p,q) by (p^*,q^*), the upshot being the symmetry of the boundary conditions.

Proof. Since $y_j(1-x,\bar\lambda,p^*,q^*)$ solves $L_{p,q}y_j = \lambda y_j$, $j = 1,2,3$, then

$$\begin{pmatrix} \bar y_1(1-x,\bar\lambda,p^*,q^*) \\ \bar y_2(1-x,\bar\lambda,p^*,q^*) \\ \bar y_3(1-x,\bar\lambda,p^*,q^*) \end{pmatrix} = N \begin{pmatrix} y_1(x,\lambda,p,q) \\ y_2(x,\lambda,p,q) \\ y_3(x,\lambda,p,q) \end{pmatrix}, \tag{4.10}$$

where N is a 3×3 matrix with complex coefficients independent of x.

At $x = 1$, (4.10) gives

$$N \begin{pmatrix} y_1 & -y_1' & y_1'' \\ y_2 & -y_2' & y_2'' \\ y_3 & -y_3' & y_3'' \end{pmatrix}(1,\lambda,p,q) = Identity.$$

Thus $\det N = -1$,

$$N = \begin{pmatrix} y_3''y_2' - y_3'y_2'' & y_1''y_3' - y_1'y_3'' & y_2''y_1' - y_2'y_1'' \\ y_3''y_2 - y_3y_2'' & y_1''y_3 - y_1y_3'' & y_2''y_1 - y_2y_1'' \\ y_3'y_2 - y_3y_2' & y_1'y_3 - y_1y_3' & y_2'y_1 - y_2y_1' \end{pmatrix}(1,\lambda,p,q)$$

and

$$N = \begin{pmatrix} \bar y_3'' & -\bar y_2'' & \bar y_1'' \\ \bar y_3' & -\bar y_2' & \bar y_1' \\ \bar y_3 & -\bar y_2 & \bar y_1 \end{pmatrix}(1,\bar\lambda,p,q) \tag{4.11}$$

using (3.1) and $q(0) = 0$. The lemma follows from (4.11) together with (4.10) and its derivatives evaluated at $x = 0$. □

Next we consider the $DP^\pm_*$ boundary conditions

$$(DP^\pm_*) \quad \begin{cases} z(0) = 0, \\ z'(1) = \pm z'(0), \\ z''(1) = 0. \end{cases}$$

Using the map $(p,q) \mapsto (p^*,q^*)$, Lemma 4.2 and section 2, it is easy to see that the increasing sequence $\left(\nu_j^\pm(p,q)\right)_{j\in\mathbf{Z}}$ of eigenvalues of $L_{p,q}$ with the $DL^\pm_*$ boundary conditions are the roots of $y_3''(1,\lambda,p,q) \pm \bar y_3''(1,\lambda,p,q)$. Lemma 3.4 holds for the

$\nu_j^{\pm}(p,q)$'s with the self-evident changes of notation. In particular, $\nu_j^{\pm}(p,q)$ is defined. Furthermore,

$$\nu_j^{\pm}(p,q) = \mu_j^{\pm}(p^*,q^*) \quad \forall\, j \in \mathbf{Z}.$$

Each $\nu_j^{\pm}(p,q)$ is of multiplicity one if $y_3''(1,\nu_j^{\pm}(p,q),p,q) \neq 0$.

Then, define for a.e. $x \in [0,1]$ and for $\lambda \in \mathbf{C}$,

$$z(x,\lambda,p,q) = y_3''(1,\lambda,p,q)y_2(x,\lambda,p,q) - y_2''(1,\lambda,p,q)y_3(x,\lambda,p,q). \tag{4.12}$$

For every $j \in \mathbf{Z}$, $z(\cdot,\nu_j^{\pm}(p,q),p,q)$ is the eigenfunction associated with the eigenvalue $\nu_j^{\pm}(p,q)$ if $y_3''(1,\nu_j^{\pm}(p,q),p,q) \neq 0$.

The $DP_*^{\pm}$ boundary conditions will be considered only for the operator L_{p^*,q^*}. Then observe that

$$y_3''(1,\lambda,p^*,q^*) \neq 0 \Longleftrightarrow y_1(1,\lambda,p,q) \neq 0 \quad \forall \lambda \in \mathbf{C}.$$

Of course we can replace above the statement $\forall\lambda \in \mathbf{C}$ with $\forall\lambda = \mu_j^{\pm}(p,q),\ j \in \mathbf{Z}$. All these conditions are also equivalent to the basic one $\mu^+(p,q) \cap \mu^-(p,q) = \phi$.

Now fix $(p,q) \in L_R^2 \times H_R^1$ and $(\tilde{p},\tilde{q}) \in L_R^2 \times H_R^1$ with $q(0) = \tilde{q}(0) = 0$. Suppose

$$\mu^+(p,q) \cap \mu^-(p,q) = \phi. \tag{4.13}$$

Also suppose that

$$\mu_j^+(p,q) = \mu_j^+(\tilde{p},\tilde{q}) \quad \text{and} \quad \mu_j^-(p,q) = \mu_j^-(\tilde{p},\tilde{q}). \tag{4.14}$$

For all $\lambda \in \mathbf{C}$, define

$$f(\lambda) = \frac{\left[\dfrac{y(x,\lambda,p,q)}{y_1(1,\lambda,p,q)} - \dfrac{y(x,\lambda,\tilde{p},\tilde{q})}{y_1(1,\lambda,\tilde{p},\tilde{q})}\right]\left[\dfrac{z(1-x,\lambda,p^*,q^*)}{y_3''(1,\lambda,p^*,q^*)} - \dfrac{z(1-x,\lambda,\tilde{p}^*,\tilde{q}^*)}{y_3''(1,\lambda,\tilde{p}^*,\tilde{q}^*)}\right]}{Y_1(1,\lambda,p,q)Z_1(1,\lambda,p,q)}. \tag{4.15}$$

The function $f(\lambda)$ is meromorphic and has simple poles at $\mu_j^+(p,q)$ and $\mu_j^-(p,q)$, $j \in \mathbf{Z}$.

For $\lambda = \mu_j^{\pm}(p,q) = \nu_j^{\pm}(p^*,q^*)$, $y(x,\lambda,p,q)$, and $\overline{z(1-x,\lambda,p^*,q^*)}$ are eigenfunctions of $L_{p,q}$ for the $DL^{\pm}$ boundary conditions. Then for some complex number $c_k^{\pm}(p,q)$,

$$\frac{y(x,\mu_j^{\pm}(p,q),p,q)}{y_1(1,\mu_j^{\pm}(p,q),p,q)} = c_j^{\pm}(p,q)\,\frac{\overline{z(1-x,\nu_j^{\pm}(p^*,q^*),p^*,q^*)}}{\overline{y_3''(1,\nu_j^{\pm}(p^*,q^*),p^*,q^*)}}. \tag{4.16}$$

Differentiate (4.16) with respect to x, then take $x = 1$, use $y'(1,\mu_j^{\pm}(p,q),p,q) = -\overline{y_1(1,\mu_j^{\pm}(p,q),p,q)}$ and $z'(0,\nu_j^{\pm}(p^*,q^*),p^*,q^*) = y_3''(1,\nu_j^{\pm}(p^*,q^*),p^*,q^*)$ to obtain

$$c_j^{\pm}(p,q) = \frac{\overline{y_1(1,\mu_j^{\pm}(p,q),p,q)}}{y_1(1,\mu_j^{\pm}(p,q),p,q)}. \tag{4.17}$$

Then (4.17) gives

$$c_j^+(p,q) = 1, \qquad c_j^-(p,q) = -1$$

$\forall j \in \mathbf{Z}$.

The computation of the residues of f is as follows. From (4.16)

$$\begin{aligned} &Res\left(f, \mu_j^+(p,q)\right) \\ &\quad = \frac{1}{(Y_1\dot{Z}_1)(1,\mu_j^+(p,q),p,q)}\left|\frac{y(x,\mu_j^+(p,q),p,q)}{y_1(1,\mu_j^+(p,q),p,q)} - \frac{y(x,\mu_j^+(\tilde{p},\tilde{q}),\tilde{p},\tilde{q})}{y_1(1,\mu_j^+(\tilde{p},\tilde{q}),\tilde{p},\tilde{q})}\right|^2 \end{aligned} \tag{4.18}$$

and

$$\begin{aligned} &Res\left(f, \mu_j^-(p,q)\right) \\ &\quad = -\frac{1}{(Z_1\dot{Y}_1)(1,\mu_j^-(p,q),p,q)}\left|\frac{y(x,\mu_j^-(p,q),p,q)}{y_1(1,\mu_j^-(p,q),p,q)} - \frac{y(x,\mu_j^-(\tilde{p},\tilde{q}),\tilde{p},\tilde{q})}{y_1(1,\mu_j^-(\tilde{p},\tilde{q}),\tilde{p},\tilde{q})}\right|^2 \end{aligned} \tag{4.19}$$

$\forall j \in \mathbf{Z}$.

Therefore (4.18), (4.19), and Lemma 4.1 give

$$\forall\, j \in \mathbf{Z}, \quad Res\left(f, \mu^+(p,q)\right) \le 0, \text{ and } Res\left(f, \mu^-(p,q)\right) \le 0. \tag{4.20}$$

Besides, the numerator $n(\lambda)$ of f is bounded from above using

$$\frac{y(x,\lambda,p,q)}{y_1(1,\lambda,p,q)} = \frac{y(x,\lambda,0,0)}{y_1(1,\lambda,0,0)} + O\left(\frac{e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x}}{|k|^2}\right) \tag{4.21}$$

and

$$\begin{aligned} &\frac{z(x,\lambda,p,q)}{y_3''(1,\lambda,p,q)} + 2q(1)\frac{y_3(x,\lambda,0,0) - y_2(x,\lambda,0,0)}{y_3''(x,\lambda,0,0)} \\ &\qquad = \frac{z(x,\lambda,0,0)}{y_3''(1,\lambda,0,0)} + O\left(\frac{e^{\left(|\Im m\frac{k}{2}|+|\Im m\frac{\omega k}{2}|+|\Im m\frac{\omega^2 k}{2}|\right)x}}{|k|^2}\right). \end{aligned} \tag{4.22}$$

Identities (4.21) and (4.22) are derived using (2.24) and (2.25).

Observe that $y_1(1,\lambda,0,0)$ and $y_3''(1,\lambda,0,0)$ never vanish on $\mathbf{C}$.

From (4.21) and (4.22), if $q(1) = \tilde{q}(1)$ we have $|n(\lambda)| = O(\frac{\Xi(x,\lambda)}{|k|^2})O(\frac{\Xi(1-x,\lambda)}{|k|^2})$; thus

$$|n(\lambda)| = O\left(\frac{\Xi(1,\lambda)}{|k|^4}\right). \tag{4.23}$$

The denominator $d(\lambda)$ of $f(\lambda)$ is easily bounded from below as follows.

Combining (3.13) and (4.10) we have

$$|Y_1(1,\lambda,p,q)| \ge \frac{1}{192}\Xi(1,\lambda) \text{ for } k \in \Omega^- \text{ with } |k| > 192e^{\|(p,q)\|_{L^2_R\times H^1_R}}$$

and

$$|Z_1(1,\lambda,p,q)| \ge \frac{1}{96}\Xi(1,\lambda) \text{ for } k \in \Omega^+ \text{ with } |k| > 96e^{\|(p,q)\|_{L^2_R\times H^1_R}}.$$

Therefore,

$$|d(\lambda)| > C\Xi(1,\lambda) \text{ for } k \in \Omega^+ \cap \Omega^- \text{ with } |k| > 192e^{\|(p,q)\|_{L^2_R \times H^1_R}}, \tag{4.24}$$

where C is a fixed numerical constant.

Finally (4.23) and (4.24) prove

$$f(\lambda) = O\left(\frac{1}{|k|^4}\right) \text{ for } k \in \Omega^+ \cap \Omega^- \text{ with } |k| > 192e^{\|(p,q)\|_{L^2_R \times H^1_R}}. \tag{4.25}$$

We now conclude that the sum of the residues is zero using the classical integration contour or, equivalently (see [P-Tr, Lemma 3.2]), by proving that $\sup_{|\lambda|=r_n} |f(\lambda)| = o(1/r_n)$ as $n \to +\infty$ for an unbounded sequence of positive real number r_n. From (4.25) this condition is verified by the function f defined in (4.15) with $r_n = (n+\frac{1}{2})^3\pi^3$, n sufficiently large.

Consequently, using (4.20) all residues of f are zero. In particular, there exists a function ϕ satisfying $L_{p,q}\phi = \mu_0^+(p,q)\phi$ and $L_{\tilde{p},\tilde{q}}\phi = \mu_0^+(p,q)\phi$. Then

$$2i(q-\tilde{q})\phi' + (i(q-\tilde{q})' + (p-\tilde{p}))\,\phi = 0. \tag{4.26}$$

Multiply (4.26) by $\overline{\phi}$ and subtract ϕ multiplied by the complex conjugate of (4.26) to have

$$\left((q-\tilde{q})|\phi|^2\right)' = 0. \tag{4.27}$$

Similarly, multiply (4.26) by $\overline{\phi}$ and add to ϕ multiplied by the complex conjugate of (4.26) to obtain

$$(p-\tilde{p})|\phi|^2 = 0. \tag{4.28}$$

Using (2.2), $q(0) = \tilde{q}(0) = 0$ and the continuity of q and $\tilde{q}$, the proof follows (4.28) and (4.27). □

5. Proof of Theorem 1.2. We now prove the asymptotic behavior of $\mu_j^+(p,q)$ given in Theorem 1.2. The proof of the asymptotic behavior of $\mu_j^-(p,q)$ is similar and we will omit it.

From (2.14), (2.16), (3.11), we have

$$\begin{aligned} 0 &= Z_1(1, \mu_j^+(p,q), p, q) \\ &= Z_1(1, \mu_j^+(p,q), 0, 0) + \Im m\, c_j^1(1, \mu_j^+(p,q), p, q) + O\left(\frac{\Xi(1,k_j)}{|k_j|^2}\right), \end{aligned} \tag{5.1}$$

where $\mu_j^+(p,q) = k_j^3$. Using (2.15)

$$\begin{aligned} &c_j^1(1, \mu_j^+(p,q), p, q) \\ &= -2\int_0^1 y_1(t, \mu_j^+(p,q), 0, 0) y_3'(1-t, \mu_j^+(p,q), 0, 0) q(t)\, dt + O\left(\frac{\Xi(1,k_j)}{|k_j|^2}\right). \end{aligned} \tag{5.2}$$

Besides, for $p = q \equiv 0$, $\lambda \in \mathbf{C}$, and $x, t \in [0,1]$, it is easy to check using (2.3) that

$$3y_1(t)y_3'(x-t) = y_2(x) + y_2\left(x - t(1-\omega)\right) + y_2\left(x - t(1-\omega^2)\right). \tag{5.3}$$

Integrating by part (5.2) and using $q \in H^1$ and (5.3) at $x = 1$, we obtain

$$c_j^1(1, \mu_j^+(p,q), p, q) = -\frac{2}{3}[q] y_2(1, \mu_j^+(p,q), p, q) + O\left(\frac{\Xi(1,k_j)}{|k_j|^2}\right).$$

Then

$$\begin{aligned} &\Im m c_j^1(1, \mu_j^+(p,q), p, q) \\ &\quad = \frac{2}{9k_j}[q]\left(\cos k_j + e^{-\frac{\sqrt{3}}{2}k_j}\cos\left(\frac{k_j}{2} + \frac{2\pi}{3}\right) + e^{\frac{\sqrt{3}}{2}k_j}\cos\left(\frac{k_j}{2} - \frac{2\pi}{3}\right)\right) \\ &\quad\quad + O\left(\frac{\Xi(1,k_j)}{|k_j|^2}\right). \end{aligned} \tag{5.4}$$

Besides, using Lemma 2.1

$$\begin{aligned} Z_1(1, \mu_j^+(p,q), 0, 0) &= -\frac{4}{3}\sin\frac{k_j}{2}\left|\sin\omega\frac{k_j}{2}\right|^2 \\ &\sim -\frac{1}{3}\sin\frac{k_j}{2}e^{\frac{\sqrt{3}}{2}|k_j|}. \end{aligned} \tag{5.5}$$

The counting lemma gives $k_j = 2(\pi j + \delta_j)$ with $|\delta_j| < \pi$.

Therefore, using (5.4), (5.5), and

$$O\left(\Xi(1,k_j)\right) = O\left(e^{\frac{\sqrt{3}}{2}|k_j|}\right),$$

we obtain from (5.1)

$$\sin\delta_j = \frac{2}{3k_j}[q]\cos\left(\pm\delta_j - \frac{2\pi}{3}\right) + O\left(\frac{1}{|k_j|^2}\right). \tag{5.6}$$

The irrelevant sign before δ_j in (5.6) depends on the sign of k_j. Then $\delta_j \to 0$ as $j \to +\infty$. Furthermore,

$$\delta_j = -\frac{[q]}{3k_j} + O\left(\frac{1}{|k_j|^2}\right), \tag{5.7}$$

which completes the proof. □

Acknowledgments. It is a pleasure to thank J. C. Guillot for many discussions and valuable suggestions.

REFERENCES

[Am] L. Amour, *Inverse spectral theory for the AKNS systems with separated boundary conditions*, Inverse Problems, 9 (1993), pp. 507–523.

[Am-Gu] L. Amour and J.-C. Guillot, *Isospectral sets for AKNS systems with generalized periodic boundary conditions*, Geom. Funct. Anal., 6 (1996), pp. 1–27.

[Ba] V. Barcilon, *Inverse problems for a vibrating beam*, Z. Angew. Math. Phys., 27 (1976), pp. 347–358.

[Bo] G. Borg, *Eine umkehrung der Sturm–Liouvilleschen eigenwertaufgabe*, Acta Math., 78 (1946), pp. 1–96.

[Ca-Pe-Sc] L. F. Caudill, P. A. Perry, and A. W. Schueller, *Isospectral sets for fourth-order ordinary differential operators*, University of Kentucky, Lexington, KY, 1997, preprint.

[Da-Tr] B. E. J. Dahlberg and E. Trubowitz, *The inverse Sturm–Liouville problem* III, Comm. Pure Appl. Math., 37 (1984), pp. 255–267.

[Ge-Le] I. M. Gelfand and B. M. Levitan, *On the determination of a differential equation from its spectral function*, Amer. Math. Soc. Trans., 1 (1953), pp. 253–304.

[Gl] G. M. L. Gladwell, *Inverse Problems in Vibration*, Martinus Nijhoff Publishers, Boston, 1986.

[Le] Z. I. Leibenzon, *An inverse problem of spectral analysis of ordinary differential operators of higher order*, Trans. Moscow Math. Soc., 34 (1966), pp. 78–163.

[Le-Ga] B. M. Levitan and M. G. Gasymov, *The determination of a differential equation from two spectra*, Uspekhi Mat. Nauk, 19 (1964), pp. 3–63.

[Ma] V. A. Marchenko, *Some problems in the theory of second-order differential operators*, Dokl. Akad. Nauk SSSR, 72 (1950), pp. 457–560.

[McKe] H. P. McKean, *Boussinesq's equation on the circle*, Comm. Pure Appl. Math., 34 (1981), pp. 599–691.

[McLa1] J. McLaughlin, *An inverse eigenvalue problem of order four—An infinite case*, SIAM J. Math. Anal., 9 (1978), pp. 395–413.

[McLa2] J. McLaughlin, *Analytical methods for recovering coefficients in differential equations from spectral data*, SIAM Rev., 28 (1986), pp. 53–72.

[P-Tr] J. Pöschel and E. Trubowitz, *Inverse Spectral Theory*, Academic Press, New York, 1987.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1029–1056

ORTHOGONAL POLYNOMIALS AND THE CONSTRUCTION OF PIECEWISE POLYNOMIAL SMOOTH WAVELETS*

G. C. DONOVAN†, J. S. GERONIMO‡, AND D. P. HARDIN§

Abstract. Orthogonal polynomials are used to construct families of C^0 and C^1 orthogonal, compactly supported spline multiwavelets. These families are indexed by an integer which represents the order of approximation. We indicate how to obtain from these multiwavelet bases for $L^2[0,1]$ and present a C^2 example.

Key words. orthogonal polynomials, multiwavelets, splines

AMS subject classification. 41A15

PII. S0036141096313112

1. Introduction.

Wavelet bases [2] for $L^2(\mathbf{R})$ have the nice property that once one of the basis functions is known the rest may be obtained by dilation and integer translation of this function. In this case the basis has one generator. Multiwavelets [1], [5], [9], [11], [13], [14], [19] are similar to wavelets except that the basis is obtained by the dilation and integer translation of several functions instead of just one. The construction of most wavelets and multiwavelets is based on multiresolution analysis (MRA) [17], [16]. Let $\phi^0, \dots, \phi^r$ be compactly supported L^2-functions, and suppose that $V_0 = \mathrm{cl}_{L^2} \operatorname{span}\{\phi^i(\cdot - j)\colon\ i = 0, 1, \dots, r,\ j \in \mathbf{Z}\}$. Then V_0 is called a *finitely generated shift invariant* (FSI) *space*. Let $(V_p)_{p\in\mathbf{Z}}$ be given by $V_p = \{\phi(2^p\cdot)\colon\ \phi \in V_0\}$. Each space V_p may be thought of as approximating L^2 at a different resolution depending on the value of p. The sequence (V_p) is called a *multiresolution analysis* generated by $\phi^0, \dots, \phi^r$ if (a) the spaces are nested, $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$, and (b) the generators $\phi^0, \dots, \phi^r$ and their integer translates form a Riesz basis for V_0. Because of (a) and (b), we can write

$$V_{j+1} = V_j \oplus W_j \quad \forall j \in \mathbf{Z}. \tag{1.1}$$

The space W_0 is called the *wavelet space*, and if $\psi^0, \dots, \psi^r$ generate a shift-invariant basis for W_0, then these functions are called *wavelet functions*. If, in addition, $\phi^0, \dots, \phi^r$ and their integer translates form an orthogonal basis for V_0, then (V_p) is called an *orthogonal MRA*. It has been shown in Lemma 2.1 of [5] that we can always assume that ϕ^j, $j = 0, \dots, r$ can be chosen so that they are *minimally supported* on $[-1, 1]$; i.e., each scaling function has support in $[-1, 1]$, and the nonzero restrictions of the scaling functions and their integer translates to $[0, 1]$ are linearly independent. Set $\Phi = (\phi^0, \dots, \phi^r)^*$. Then (a) and (b) imply that Φ satisfies the matrix dilation equation

$$\Phi(t) = \sum C_i \Phi(2t - i). \tag{1.2}$$

*Received by the editors December 4, 1996; accepted for publication (in revised form) August 3, 1998; published electronically August 16, 1999. This research was supported in part by the Institute for Mathematics and its Applications with funds provided by the NSF.
http://www.siam.org/journals/sima/30-5/31311.html

†Department of Mathematics, Princeton University, Princeton, NJ 08544 (donovan@math.princeton.edu). This author was supported by an NSF postdoctoral fellowship.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (geronimo@math.gatech.edu). This author was supported by an NSF grant.

§Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (hardin@math.vanderbilt.edu). This author was supported by an NSF grant.

In order to obtain orthogonal MRAs it is useful to divide $V_0|_{[0,1]}$ into the following subspaces:

- $A_0 = \operatorname{span}\{\phi \in \Phi\colon \ \operatorname{supp}\phi \subset [0,1]\} = \{g \in V_0\colon \ \operatorname{supp} g \subset [0,1]\}$,
- $C_0 = \operatorname{span}\{\phi(\cdot)\chi_{[0,1]}\colon \ \phi \in \Phi\} \ominus A_0$,
- $C_1 = \operatorname{span}\{\phi(\cdot - 1)\chi_{[0,1]}\colon \ \phi \in \Phi\} \ominus A_0$.

Since the functions in A_0 are orthogonal to their integer translates and by Gram–Schmidt may be made mutually orthogonal, it is not difficult to see [5] the following.

THEOREM 1.1. *(V_p) is orthogonal iff $C_0 \perp C_1$.*

In general it is not possible to obtain an orthogonal basis for V_0 by taking finite linear combinations of the original basis functions that generate it. However, we can modify V_0 by adding appropriate functions so that $C_0 \perp C_1$. The notion of intertwining MRAs arises from the observation that some functions can be moved from one level of an MRA to another without destroying the defining properties of an MRA. Suppose that (V_p) is an MRA as defined above and that W is an FSI subspace of V_1 with Riesz basis $\{\phi^{r+1}(\cdot - j)\colon \ j \in \mathbf{Z}\}$ for some compactly supported ϕ^{r+1}. With this we can generate a new space $\tilde{V}_0 = V_0 + W$ together with associated dilation spaces $\tilde{V}_p = \{\phi(2^p\cdot)\colon \ \phi \in \tilde{V}_0\}$. Then $(\tilde{V}_p)$ is also a multiresolution analysis. Indeed, it is clear that $V_0 \subset \tilde{V}_0$. It is also the case that $\tilde{V}_0 \subset V_1$, since both V_0 and W are subspaces of V_1. From these two inclusions and the definition of the dilated spaces, it follows that

$$\cdots \subset V_{-1} \subset \tilde{V}_{-1} \subset V_0 \subset \tilde{V}_0 \subset V_1 \subset \tilde{V}_1 \cdots \tag{1.3}$$

and hence

$$\cdots \subset \tilde{V}_{-1} \subset \tilde{V}_0 \subset \tilde{V}_1 \subset \cdots.$$

It is easy to see how the other conditions necessary for $(\tilde{V}_p)$ to be an MRA also follow from (1.3). Of course this process can be repeated if more functions are needed. Let $A_1 = \{\phi \in V_1\colon \ \operatorname{supp}\phi \subseteq [0,1]\}$. If $W \subset A_1 \ominus A_0$ such that $(I - P_W)C_0 \perp (I - P_W)C_1$, where P_W is the orthogonal projection onto W, then the intertwining MRA will be an orthogonal MRA. In Theorem 1 of [5] the following theorem was proved.

THEOREM 1.2. *If (V_p) is a multiresolution analysis generated by compactly supported scaling functions, then there is some pair of integers (q, n) and some orthogonal multiresolution analysis $(\tilde{V}_p)$ such that*

$$V_q \subset \tilde{V}_0 \subset V_{q+n}.$$

We call $(\tilde{V}_p)$ an *intertwining MRA* and the theorem implies in particular that for those MRAs generated by splines there exist orthogonal intertwining MRAs also generated by splines. The main results of this paper give explicit constructions of orthogonal intertwining MRAs for C^0 and C^1 spline MRAs with various orders of approximation. In this paper we will use the notation

$$f_{i,j} = 2^{i/2} f(2^i \cdot - j) \tag{1.4}$$

and $P_{\{f_1, f_2, \ldots, f_k\}}$, $f_i \in L^2(\mathbf{R})$, will denote the orthogonal projection onto the subspace spanned by $f_1, \ldots, f_k$. In section 2 we make precise which MRAs we will be studying and examine the spaces A_0, C_0, and C_1 associated with these MRAs. It is here where orthogonal polynomials play a role since the A_0 spaces for the MRAs we will be considering will be spanned by subclasses of ultraspherical polynomials (see also [18], [15], and [10]). Furthermore, we introduce the functions that will span W, which we

will "borrow" in order to construct an orthogonal intertwining MRA. These functions are chosen so as to be smooth as possible and symmetric or antisymmetric with respect to $x = \frac{1}{2}$. In section 3 we derive various properties of these functions needed to carry out the construction. In section 4 we give a description of how to construct wavelets from scaling functions. This explicit computation of the multiwavelets functions in terms of the scaling functions allows us to preserve symmetry or antisymmetry. Also indicated in this section is how to construct multiwavelet bases for $L^2[0,1]$. In section 5 we construct families of compactly supported continuous orthogonal scaling functions having an axis of symmetry. Each family is indexed by an integer which represents the order of approximation. In section 6 families of C^1 scaling functions and wavelets are constructed. Finally, in section 7, a C^2 example is given.

2. Piecewise polynomial MRA. The MRAs we will study are those associated with piecewise polynomial splines. Let S_k^n be the space of polynomial splines of degree n with C^k knots at the integers, and set $V_0^{n,k} = S_k^n \cap L^2(\mathbf{R})$. With $V_p^{n,k}$ as above it is known that these spaces form a multiresolution analysis [1], [5], [11], [19]; however, with the exception of S_{-1}^n, which were studied by Alpert, it is not possible to find compactly supported, orthogonal generators for these spaces. We will consider S_k^n when $n > 2k+1$.

In order to construct orthogonal intertwining MRAs we examine the spaces $A_0^{n,k}$, $C_0^{n,k}$, and $C_1^{n,k}$, associated with $V_0^{n,k}$ as described above, as well as $A_1^{n,k}$, defined by analogy with $A_0^{n,k}$ to be $A_1^{n,k} = \{g \in V_1\colon \operatorname{supp} g \subset [0,1]\}$. Our goal is to find a space $W \subset A_1^{n,k} \ominus A_0^{n,k}$ so that

$$(I - P_W)C_0^{n,k} \perp (I - P_W)C_1^{n,k}. \tag{2.1}$$

We begin by describing a convenient basis for $A_0^{n,k}$. Note that from the description of $A_0^{n,k}$, $g \in A_0^{n,k}$ iff $g(t) = t^{k+1}(1-t)^{k+1}q(t)$ for some polynomial q of degree at most $n-2k-2$. Thus the linear dimension of $A_0^{n,k}$ is $n-2k-1$. If $\{t^{k+1}(1-t)^{k+1}q_{j-2k-2}\colon j = 2k+2,\dots,n\}$ forms an orthogonal basis for $A_0^{n,k}$, then

$$\int_0^1 t^{2k+2}(1-t)^{2k+2}q_{i-2k-2}(t)q_{j-2k-2}(t)\,dt = c_i\delta_{i,j},$$

and we see that the q_i's are orthogonal with respect to the measure $t^{2k+2}(1-t)^{2k+2}\,dt$ on $[0,1]$. The monic ultraspherical polynomials p_i^λ are monic polynomials of degree $i = 0,1,\dots,$ which are orthogonal on $[-1,1]$ with respect to the measure $(1-t^2)^{\lambda-\frac{1}{2}}dt$ (Szegö [20]). From the above equation, we can choose $q_i(t) = p_i^{2k+\frac{5}{2}}(2t-1)$, $i = 0,\dots,n-2k-2$. This leads to ([7], [8]) the following lemma.

LEMMA 2.1. *A basis for $A_0^{n,k}$ is $\{t^{k+1}(1-t)^{k+1}p_i^{2k+\frac{5}{2}}(2t-1)\colon i = 0,\dots,n-2k-2\}$, where each $p_i^{2k+\frac{5}{2}}$ is a monic ultraspherical polynomial of degree i.*

The ultraspherical polynomials already have been used in the construction of wavelets from fractal interpolation functions (Donovan, Geronimo, and Hardin [6], Donovan [4]). For later computations we define the set $\{\phi_i^k\}$, where $\phi_i^k(t) = (1-t^2)^{k+1}p_{i-2k-2}^{2k+\frac{5}{2}}(t)$ for $i \geq 2k+2$ with $\phi_{2k+1}^k = 0$.

Some important properties of the ultraspherical polynomials that we will use later [20] follow.

(a) The Rodriguez formula:

$$(1-t^2)^m p_n^{m+\frac{1}{2}}(t) = (-1)^n \frac{(n+2m)!}{(2n+2m)!} D^n(1-t^2)^{n+m}.$$

(b) The recurrence formula:

$$p_{n+1}^{m+\frac{1}{2}}(t) = tp_n^{m+\frac{1}{2}}(t) - a_n p_{n-1}^{m+\frac{1}{2}}(t), \quad n = 1, 2, \ldots$$

with $a_n = \frac{(n+2m)n}{(2n+2m+1)(2n+2m-1)}$, $p_0(t) = 1$ and $p_1(t) = t$.

(c)

$$p_n^{m+\frac{1}{2}}(t) = \frac{1}{(n+2)(n+1)} D^2 p_{n+2}^{m-2+\frac{1}{2}}(t).$$

The polynomials also have the following useful representation in term of hypergeometric functions [20]:

$$p_n^{m+\frac{1}{2}}(t) = \frac{2^n (m+1)_n}{(n+2m+1)_n} \, {}_2F_1\left(\begin{matrix} -n & n+2m+1 \\ & m+1 \end{matrix}; \ \frac{1-t}{2}\right), \tag{2.2}$$

where formally

$${}_pF_q\left(\begin{matrix} a_1 & \ldots & a_p \\ b_1 & \ldots & b_q \end{matrix}; \ t\right) = \sum_{i=0}^{\infty} \frac{(a_1)_i \ldots (a_p)_i}{(b_1)_i \ldots (b_q)_i (1)_i} t^i$$

with $(a)_i = (a)(a+1)\ldots(a+i-1)$. Since one of the numerator parameters in the hypergeometric function in (2.2) is a negative integer, the series has finitely many nonzero terms and the result is a polynomial.

From the recurrence formula is it not difficult to see that

$$\int_{-1}^{1} p_n^{m+\frac{1}{2}}(t)^2 (1-t^2)^m dt = \frac{2(n+2m)!n!}{(2n+2m-1)!!(2n+2m+1)!!}, \tag{2.3}$$

where $m!! = m(m-2)\ldots$. Thus

$$\|\phi_n^k\|_{L^2}^2 = \frac{2(n+2k+2)!(n-2k-2)!}{(2n-1)!!(2n+1)!!}. \tag{2.4}$$

We now consider the spaces $C_i^{n,k}$, $i = 0, 1$. Since the dimension of $V_0^{n,k}|_{[0,1]}$ is $n+1$ and the dimension of $A_0^{n,k}$ is $n-2k-1$, the dimension of $C_0^{n,k}$ is $k+1$ and the same is true for $C_1^{n,k}$. For computational compatibility with the ultraspherical polynomials, we scale these spaces so that they are defined on $[-1, 1]$ instead of $[0, 1]$ and denote them as $C_i^{n,k}(\frac{\cdot+1}{2})$ and $A_0^{n,k}(\frac{\cdot+1}{2})$.

Let $r_i^k(t) = (1-t)^i(1+t)^{k+1}$, $l_i^k(t) = r_i^k(-t)$, and $P^{n,k}$ denote the orthogonal projection onto $A_0^{n,k}$. Then a basis for $C_0^{n,k}(\frac{\cdot+1}{2})$ is $r_i^{n,k} = (I - P^{n,k})r_i^k$, $i = 0, \ldots, k$, while for $C_1^{n,k}(\frac{\cdot+1}{2})$ the analogous basis is $l_i^{n,k} = (I - P^{n,k})l_i^k$. We shall refer to the families of r_i^k and l_i^k, respectively, as right and left "ramps," where right and left denote on which side of the interval $[-1, 1]$ the unprojected functions *do not* vanish $k+1$ times.

The action of the projection can be readily computed from the following inner product:

$$\langle r_i^k, \phi_n^k \rangle = \frac{2^{k+i+n+2} n!(i+k+1)!(n-k-i-2)!(n+2k+2)!}{(k-i)!(k+i+n+2)!(2n)!}, \tag{2.5}$$

where $k \geq i$ and $n \geq 2k+2$ (Gradshteyn and Ryzhik [12, p. 826]). This formula can be derived for instance by substituting (2.2) into the above inner product, integrating term by term, then using Saalschutz's formula for summing a ${}_3F_2$ hypergeometric function (Bailey [3, p. 49]). The symmetry between r_i^k and l_i^k and the fact that the ultraspherical polynomials are of definite parity imply that $\langle l_i^k, \phi_n^k \rangle = (-1)^n \langle r_i^k, \phi_n^k \rangle$.

There are several simple useful relations among the ramp functions which we now derive. The first is the obvious relation

$$r_i^{n,k} = r_i^{n-1,k} - \frac{\langle r_i^k, \phi_n^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_n^k. \tag{2.6}$$

Note that for $n \geq 2k+2$, the polynomial $r_{i+1}^{n,k}$ is uniquely determined by the order of its zeros at ± 1, its orthogonality to $A_0^{n,k}$, its degree, and its leading coefficient. The polynomial $(1-t)r_i^{n,k}$ vanishes at ± 1 the same number of times as $r_{i+1}^{n,k}$, is of degree $n+1$, and is orthogonal to $A_0^{n-1,k}$. Subtracting off ϕ_{n+1}^k times an appropriate constant then projecting out ϕ_n^k yields the following relation:

$$r_{i+1}^{n,k}(t) = (1-t)r_i^{n,k}(t) + \frac{\langle r_i^k, \phi_{n+1}^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_n^k(t) - \frac{\langle r_i^k, \phi_n^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_{n+1}^k(t). \tag{2.7}$$

Since $(1-t^2)\frac{d}{dt} r_i^k(t) = (k+1+i)r_{i+1}^k - 2ir_i^k$, an argument similar to that given above leads to

$$\begin{aligned}(1-t^2)\frac{d}{dt} r_i^{n,k}(t) = (k+1+i)r_{i+1}^{n,k} - 2ir_i^{n,k} + (n+2)\frac{\langle r_i^k, \phi_{n+1}^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_n^k(t) \\ + n\frac{\langle r_i^k, \phi_n^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_{n+1}^k(t).\end{aligned} \tag{2.8}$$

Likewise,

$$\begin{aligned}\frac{d}{dt}\big((1-t^2)r_i^{n,k}(t)\big) = (k+3+i)r_{i+1}^{n,k} - 2(i+1)r_i^{n,k} + n\frac{\langle r_i^k, \phi_{n+1}^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_n^k(t) \\ + (n+2)\frac{\langle r_i^k, \phi_n^k \rangle}{\langle \phi_n^k, \phi_n^k \rangle} \phi_{n+1}^k(t).\end{aligned} \tag{2.9}$$

Analogous formulas are easily obtained for $l_i^{n,k}$. In the multiresolution analysis given above, $V_1^{n,k}$ is a spline space with knots at the half-integers. Thus, to do the intertwining step, we will borrow from the space of splines supported on $[0,1]$ with a knot at $\frac{1}{2}$. Once again, for purposes of compatibility with the ϕ_i^k, we dilate these functions so that they are supported on $[-1,1]$ and have a knot at 0. The dilated functions will be denoted u_n^k. To construct these, we first define a sequence of spline functions $\{\bar{u}_n^k\} \in C^k(\mathbf{R})$, where

$$\bar{u}_n^k(t) = \begin{cases} 1 - |t|\, t^{n-1} + \sum_{i=1}^k \frac{(-1)^i n!!}{2^i i!\,(n-2i)!!}(1-t^2)^i & \text{if } t \in [-1,1], \\ 0 & \text{otherwise} \end{cases}$$

for odd $n \geq 2k+1$, and

$$\bar{u}_n^k(t) = \begin{cases} t - |t|\, t^{n-1} + \sum_{i=1}^k \frac{(-1)^i (n-1)!!}{2^i i!\,(n-2i-1)!!} t(1-t^2)^i & \text{if } t \in [-1,1], \\ 0 & \text{otherwise} \end{cases}$$

for even $n \geq 2k+2$. It should be noted that the parity of the function $\bar{u}_n^k$ is always the opposite of the parity of n. From this sequence, we obtain new sequences by applying the projection $I - P^{n,k}$:

$$u_{n,m}^k = (I - P^{n,k})\bar{u}_{n-m}^k.$$

The functions to be borrowed from $V_1(\frac{\cdot+1}{2})$ to give an orthogonal multiresolution analysis are obtained as appropriate linear combinations of this latter class of splines. To calculate the above projections, it is necessary to find the inner products

$$c_{k,n,j} = \int_{-1}^{1} \bar{u}_j^k(t)\phi_n^k(t)dt$$

for $n,\ j \geq 2k+1$, which are given in the following lemma.

LEMMA 2.2. *With $n, j \geq 2k+2$, if $(n-j)$ is even, then $c_{k,n,j} = 0$, and if $(n-j)$ is odd and $j > n$, then*

$$\begin{aligned} c_{k,n,j} &= \frac{-2^{\frac{n-j+4k+3}{2}}(n-2k-2)!j!(\frac{j-2k-n}{2})_{k+1}}{(n+1)_n(j-n)!!(\frac{j-2k+n-1}{2})!} \\ &\qquad \times {}_3F_2\left(\begin{matrix} -k-1 & \frac{j+2}{2} & \frac{j+1}{2} \\ \frac{j-2k-n}{2} & \frac{j-2k+n+1}{2} \end{matrix};\ 1\right), \\ &= \frac{-2^{\frac{n-j+4k+3}{2}}(\frac{n-2k}{2})_{k+1}(\frac{-2k-n-1}{2})_{k+1}(n-2k-2)!j!}{(n+1)_n(j-n)!!(\frac{j+n+1}{2})!} \\ &\qquad \times {}_3F_2\left(\begin{matrix} -k-1 & \frac{j+1}{2} & k+1 \\ & \frac{-n}{2} & \frac{n+1}{2} \end{matrix};\ 1\right), \end{aligned} \tag{2.10}$$

while if $(n-j)$ is odd and $n > j$, then

$$\begin{aligned} c_{k,n,j} &= \frac{2\,(-1)^{\frac{n-j+1}{2}+k+1}(n-2k-2)!j!(n+2k-j+1)!}{(n+1)_n(\frac{n-2k+j-1}{2})!(\frac{n-j+2k+1}{2})!} \\ &\qquad \times {}_3F_2\left(\begin{matrix} -k-1 & \frac{j+2}{2} & \frac{j+1}{2} \\ \frac{j-2k-n}{2} & \frac{j-2k+n+1}{2} \end{matrix};\ 1\right) \\ &= \frac{(-1)^{\frac{n-j+1}{2}}2^{2k+3}(\frac{n-2k}{2})_{k+1}(\frac{-2k-n-1}{2})_{k+1}\ (n-2k-2)!j!(n-j-1)!}{(n+1)_n(\frac{n+j+1}{2})!(\frac{n-j-1}{2})!} \\ &\qquad \times {}_3F_2\left(\begin{matrix} -k-1 & \frac{j+1}{2} & k+1 \\ & \frac{-n}{2} & \frac{n+1}{2} \end{matrix};\ 1\right). \end{aligned} \tag{2.11}$$

Remark. When $n = 2k+2$ the ${}_3F_2$ in the second part of (2.11) reduces to a truncated ${}_2F_1$.

Proof. The fact that $c_{k,n,j} = 0$ for $(n-j)$ even follows from the parity of $\bar{u}_j^k$ and ϕ_n^k. From property (c) above we find the $p_n^{2k+\frac{5}{2}} = \frac{n!}{(n+2k+2)!}D^{2k+2}p_{n+2k+2}^{1/2}$, where $p_n^{1/2}$ is the monic Legendre polynomial of degree n [20]. Therefore from the definition of ϕ_n^k we find that $\phi_n^k(t) = (1-t^2)^{k+1}\frac{(n-2k-2)!}{n!}D^{2k+2}p_n^{1/2}(t)$. Substitute this into the integral for $c_{m,n,j}$ and integrate by parts $2k+2$ times. Since $j \geq 2k+2$, $\bar{u}_j^k$ has $2k+1$

continuous derivatives in $(-1,1)$ and $D^i(1-t^2)^{k+1}\bar{u}_j^k = 0$ for $t = \pm 1$, $i = 0, \ldots, 2k+1$. Therefore

$$c_{k,n,j} = \frac{(n-2k-2)!}{n!}\int_{-1}^{1} D^{2k+2}[(1-t^2)^{k+1}\bar{u}_j^k(t)]p_n^{1/2}(t)dt.$$

Observe that for $i < k+1$, $D^{2k+2}(1-t^2)^{k+i+1}$ is a polynomial of degree less than $2k+2$. Consequently the orthogonality of $p_n^{1/2}$ implies that only the term $|t|t^{j-1}$ in $\bar{u}_j^k$ will be nonzero in the above integral. Using the parity of $\bar{u}_j^k$ and ϕ_n^k to integrate on $[0,1]$, then expanding $(1-t^2)^{k+1}$ and differentiating yields

$$c_{k,n,j} = -2\frac{(n-2k-2)!}{n!}\sum_{i=0}^{k+1}\binom{k+1}{i}(-1)^i\frac{(j+2i)!}{(j+2i-2k-2)!}\int_0^1 t^{j+2i-2k-2}p_n^{1/2}(t)dt.$$

With $l = j+2i-2k-2$ we find, using (2.2) with $m=0$, that

$$\begin{aligned}\int_0^1 t^l p_n^{1/2}(t)dt &= \frac{2^n n!l!}{(n+1)_n(l+1)!}{}_2F_1\left(\begin{matrix}-n & n+1\\ & l+2\end{matrix};\ \tfrac{1}{2}\right)\\ &= \frac{2^n n!l!\Gamma(\frac{l+3}{2})\Gamma(\frac{l+2}{2})}{(n+1)_n(l+1)!\Gamma(\frac{l+3+n}{2})\Gamma(\frac{l+2-n}{2})},\end{aligned}$$

where Kummer's Theorem has been used in summing the ${}_2F_1$ [3, p. 11]. If n is even, l is odd and we find

$$\frac{\Gamma(\frac{l+3+n}{2})}{\Gamma(\frac{l+3}{2})} = \frac{(\frac{l+n+3-2i}{2})_i(\frac{l+n-2i+1}{2})!}{(\frac{l+1}{2})!}$$

and

$$\frac{\Gamma(\frac{l+2}{2})}{\Gamma(\frac{l+2-n}{2})} = \frac{(\frac{l-2i+2-n}{2})_{k+1}l!!2^{\frac{j-n-l}{2}}}{(\frac{l-2i+2-n}{2})_i(j-n)!!}.$$

The above formulas and the substitutions $(j+2i)! = 2^{2i}j!(\frac{j+1}{2})_i(\frac{j+2}{2})_i$ and $(-1)^i\binom{k+1}{i} = \frac{(-k-1)_i}{(1)_i}$ yield the first line of (2.10) for n even. To obtain the second line of (2.10), use the transformation formula [3, p. 85]

$$\begin{aligned}&{}_3F_2\left(\begin{matrix}-(k+1) & a & b\\ e & f\end{matrix};\ 1\right)\\ &\quad= \frac{(e-b)_{k+1}(f-b)_{k+1}}{(e)_{k+1}(f)_{k+1}}{}_3F_2\left(\begin{matrix}-(k+1) & b & a+b-k-e-f\\ b-e-k & b-f-k\end{matrix};\ 1\right).\end{aligned}$$

To arrive at the first line of (2.11) for n odd perform manipulations similar to those described above on $\frac{\Gamma(\frac{l+3+n}{2})}{\Gamma(\frac{l+2}{2})}$ and $\frac{\Gamma(\frac{l+3}{2})}{\Gamma(\frac{l+2-n}{2})}$. The second line of (2.11) is obtained in a similar manner. □

The above formulas show that the hypergeometric functions have a fixed number of terms for fixed smoothness. Furthermore, the hypergeometric functions given in the second lines of (2.10) and (2.11) are a sum of positive terms. Any zero appearing in the denominator of the above formulas cancels with a zero in the numerator.

3. Recurrence formulas and inner products. As in the case of the ramp functions there are useful difference and differential difference equations relating the $u^k_{n,m}$'s. The spline $u^k_{n,m}$ is uniquely defined in terms of the degree of its zeros at ± 1, the $n-m-1$ continuous derivatives at zero, the values of the left and right $(n-m)$th derivatives at zero, its degree, and its orthogonality to $A_0^{n,k}$. Denoting by "rem" the remainder of integer division (i.e., $p \operatorname{rem} q \equiv q(\frac{p}{q} - \lfloor \frac{p}{q} \rfloor)$), we find from the definition of $\bar{u}^k_{n-m}$ that $t\bar{u}^k_{n-m-1} = \bar{u}^k_{n-m} + ((n-m) \operatorname{rem} 2) K_{k,n,m}(1-t^2)^{k+1}$. Furthermore, $\int_{-1}^{1} t u^k_{n-1,m}(t)\phi^k_i(t)dt = 0$ for $i \le n-2$. The recurrence formula for ϕ^k_n, the fact that $(1-t^2)^{k+1} \in A_0^{n,k}(\frac{\cdot+1}{2})$, and parity considerations now imply that for $n-m-1 \ge 2k+1$

$$u^k_{n,m} = t u^k_{n-1,m} - \epsilon_{k,n,m}\phi^k_{n-1+(m \operatorname{rem} 2)}, \tag{3.1}$$

where

$$\epsilon_{k,n,m} = \frac{c_{k,n+(m \operatorname{rem} 2),n-1-m}}{\langle \phi^k_{n-1+(m \operatorname{rem} 2)}, \phi^k_{n-1+(m \operatorname{rem} 2)} \rangle}. \tag{3.2}$$

Analogues of (2.8) and (2.9) can also be derived. To this end, note that $(1-t^2)D\bar{u}^k_{n-m} = -(n-m)\bar{u}^k_{n+1-m} + (n-m)\bar{u}^k_{n-1-m}$ and that

$$\begin{aligned} \langle (1-t^2)Du^k_{n,m}(t), \phi^k_i(t) \rangle &= \langle u^k_{n,m}(t), D\big((1-t^2)\phi^k_i(t)\big) \rangle \\ &= 0, \qquad i \le n-1 \end{aligned}$$

since $u^k_{n,m}$ is orthogonal to all functions of the form $(1-t^2)^{k+1}\pi_{n-2k-2}$, where π_{n-2k-2} is an arbitrary polynomial of degree less than or equal to $n-2k-2$. The parity of $Du^k_{n,m}$ now implies that

$$(1-t^2)Du^k_{n,m}(t) = -(n-m)u^k_{n+1,m}(t) + (n-m)u^k_{n-1,m}(t) + \delta_{k,n,m}\phi^k_{n+(m \operatorname{rem} 2)}, \tag{3.3}$$

where

$$\begin{aligned} \delta_{k,n,m} = &-(n-m)\frac{c_{k,n+(m \operatorname{rem} 2),n+1-m}}{\langle \phi^k_{n+(m \operatorname{rem} 2)}, \phi^k_{n+(m \operatorname{rem} 2)} \rangle} \\ &+ (n-1+(m \operatorname{rem} 2))\frac{c_{k,n-1+(m \operatorname{rem} 2),n-m}}{\langle \phi^k_{n-1+(m \operatorname{rem} 2)}, \phi^k_{n-1+(m \operatorname{rem} 2)} \rangle}. \end{aligned} \tag{3.4}$$

Another similar formula is

$$D\big((1-t^2)u^k_{n,m}(t)\big) = -(n-m+2)u^k_{n+1,m}(t) + (n-m)u^k_{n-1,m}(t) + \gamma_{k,n,m}\phi^k_{n+(m \operatorname{rem} 2)} \tag{3.5}$$

with

$$\begin{aligned} \gamma_{k,n,m} = &(n+1+(m \operatorname{rem} 2))\frac{c_{k,n-1+(m \operatorname{rem} 2),n-m}}{\langle \phi^k_{n-1+(m \operatorname{rem} 2)}, \phi^k_{n-1+(m \operatorname{rem} 2)} \rangle} \\ &+ (n+2-m)\frac{c_{k,n+(m \operatorname{rem} 2),n+1-m}}{\langle \phi^k_{n+(m \operatorname{rem} 2)}, \phi^k_{n+(m \operatorname{rem} 2)} \rangle}. \end{aligned} \tag{3.6}$$

One final useful formula is

$$u^k_{n,n-m} = u^k_{n-2,n-m-2} - \frac{c_{k,n-1,n-m+(m \operatorname{rem} 2)}}{\langle \phi^k_{n-1+(m \operatorname{rem} 2)}, \phi^k_{n-1+(m \operatorname{rem} 2)} \rangle}\phi^k_{n-1+(m \operatorname{rem} 2)}. \tag{3.7}$$

Note that the above formulas hold for $k \geq 0$ and $n - m \geq 2k + 3$. In order to find appropriate linear combinations of $\{u_{n,m}^k\}$ that solve (2.1) we need to compute the inner products of the various functions defined above. We begin with the following lemma.

LEMMA 3.1. *Given $r_i^{n,k}$ and $l_j^{n,k}$, $i, j \leq k$, as above,*

$$\langle r_i^{n,k}, l_j^{n,k}\rangle = \frac{(-1)^{n+1}2^{2k+i+j+3}(k+i+1)!(k+j+1)!(n-k-i-1)!(n-k-j-2)!(n+2k+3)!}{(k-i)!(k-j)!(k+n+j+3)!(k+n+i+2)!(n-2k-2)!} \times {}_3F_2\left(\begin{matrix} -(k-j) & 2k+i+j+4 & 1 \\ -(n-k-j-2) & k+n+j+4 \end{matrix}; 1\right). \tag{3.8}$$

Proof. Multiply (2.8) by $l_j^{n,k}$ and integrate by parts to find

$$-\langle r_i^{n,k}, D\big((1-t^2)l_j^{n,k}\big)\rangle = (k+1+i)\langle r_{i+1}^{n,k}, l_j^{n,k}\rangle - 2i\langle r_i^{n,k}, l_j^{n,k}\rangle + (n+2)\frac{\langle r_i^k, \phi_{n+1}^k\rangle\langle l_j^k, \phi_n^k(t)\rangle}{\langle \phi_n^k, \phi_n^k\rangle},$$

where the fact that $\langle l_j^{n,k}, \phi_{n+1}^k\rangle = \langle l_j^k, \phi_{n+1}^k\rangle$ has been used. Now eliminate $D\big((1-t^2)l_j^{n,k}\big)$ using the analogue of (2.9) for $l_j^{n,k}$ and collect terms to get

$$0 = (k+i+1)\langle r_{i+1}^{n,k}, l_j^{n,k}\rangle + 2(j+1-i)\langle r_i^{n,k}, l_j^{n,k}\rangle - (k+j+3)\langle r_i^{n,k}, l_{j+1}^{n,k}\rangle + n\frac{\langle r_i^k, \phi_n^k\rangle\langle l_j^k, \phi_{n+1}^k\rangle}{\langle \phi_n^k, \phi_n^k\rangle} + (n+2)\frac{\langle r_i^k, \phi_{n+1}^k\rangle\langle l_j^k, \phi_n^k\rangle}{\langle \phi_n^k, \phi_n^k\rangle}.$$

Multiply (2.7) by $l_j^{n,k}$ and integrate to obtain

$$\langle r_{i+1}^{n,k}, l_j^{n,k}\rangle = -\langle r_i^{n,k}, l_{j+1}^{n,k}\rangle + 2\langle r_i^{n,k}, l_j^{n,k}\rangle + \frac{\langle r_i^k, \phi_{n+1}^k\rangle\langle l_j^k, \phi_n^k\rangle}{\langle \phi_n^k, \phi_n^k\rangle} + \frac{\langle r_i^k, \phi_n^k\rangle\langle l_j^k, \phi_{n+1}^k\rangle}{\langle \phi_n^k, \phi_n^k\rangle}.$$

Using the above two equations to eliminate $\langle r_{i+1}^{n,k}, l_j^{n,k}\rangle$, solving for $\langle r_i^{n,k}, l_j^{n,k}\rangle$, and then using (2.5) gives

$$\langle r_i^{n,k}, l_j^{n,k}\rangle = \frac{(2k+i+j+3)}{2(k+j+2)}\langle r_i^{n,k}, l_{j+1}^{n,k}\rangle + \frac{(-1)^{n+1}2^{2k+i+j+3}(n+2k+3)!(k+j+1)!(k+i+1)!(n-k-j-2)!(n-k-i-1)!}{(k-j)!(k-i)!(k+j+n+3)!(k+i+n+2)!(n-2k-2)!}.$$

Iterating this formula from $j+1$ to k and utilizing the fact that l_{k+1}^k is in $A_0^{n,k}$ yields the formula

$$\langle r_i^{n,k}, l_j^{n,k}\rangle = \frac{(-1)^{n+1}2^{2k+i+j+3}(k+i+1)!(k+j+1)!(n-k-i-1)!(n+2k+3)!}{(k-i)!(k+n+i+2)!(n-2k-2)!} \times \sum_{m=0}^{k-j}\frac{(n-k-m-j-2)!(2k+i+j+4)_m}{(k-m-j)!(n+k+m+j+3)!}.$$

Finally, with the substitutions $(n+k+m+j+3)! = (n+k+j+3)!(n+k+j+4)_m$, $(a-j)! = (-1)^j\frac{a!}{(-a)_j}$, and $a \in \{k-m, n-k-m-2\}$ we obtain (3.8). ∎

An analogous argument gives

$$\begin{aligned}2(i+j+1)\langle r_i^{n,k}, r_j^{n,k}\rangle &= (k+i+1)\langle r_{i+1}^{n,k}, r_{j+1}^{n,k}\rangle + (k+j+3)\langle r_i^{n,k}, r_{j+1}^{n,k}\rangle \\ &\quad + (n+k+1+i)\frac{\langle r_i^{n,k}, \phi_n^k\rangle\langle r_j^{n,k}, \phi_{n+1}^k\rangle}{\langle \phi_n^k, \phi_n^k\rangle} \\ &\quad + (n-k-i+1)\frac{\langle r_i^{n,k}, \phi_{n+1}^k\rangle\langle r_j^{n,k}, \phi_n^k\rangle}{\langle \phi_n^k, \phi_n^k\rangle}.\end{aligned} \tag{3.9}$$

Since $\langle r_i^{n,k}, u_{n,m}^k\rangle = \langle r_i^k, u_{n,m}^k\rangle$ the following recurrence formula is easily obtained from (3.1) using the relation $r_i^k(t) = (1-t)r_{i-1}^k$:

$$\langle r_i^{n,k}, u_{n,m}^k\rangle = \langle r_i^{n-1,k}, u_{n-1,m}^k\rangle - \langle r_{i+1}^{n,k}, u_{n,m}^k\rangle + \epsilon_{k,n,m}\langle r_i^k, \phi_{n+(m \text{ rem } 2)}^k\rangle. \tag{3.10}$$

To obtain the next formula we will need the following lemma.

LEMMA 3.2. *Formally*

$$\begin{aligned}{}_3F_2\left(\begin{matrix}a & b & c\\ d & e\end{matrix};\ 1\right) &= \frac{c(e-a)}{de}{}_3F_2\left(\begin{matrix}a & b+1 & c+1\\ d+1 & e+1\end{matrix};\ 1\right) \\ &\quad + \frac{d-c}{d}{}_3F_2\left(\begin{matrix}a & b+1 & c\\ d+1 & e\end{matrix};\ 1\right).\end{aligned}$$

Proof. The proof will be based on the contiguous relations for ${}_3F_2$ hypergeometric functions found in Wilson's paper [21]. Although the proof will be formal in practice we will apply the result when one of the numerator parameters is a negative integer, in which case the series has only a finite number of nonzero terms.

Term by term subtraction yields

$${}_3F_2\left(\begin{matrix}a & b & c\\ d & e\end{matrix};\ 1\right) = \frac{bc}{de}{}_3F_2\left(\begin{matrix}a & b+1 & c+1\\ d+1 & e+1\end{matrix};\ 1\right) + {}_3F_2\left(\begin{matrix}a-1 & b & c\\ d & e\end{matrix};\ 1\right),$$

and equation 8 in [21] implies that

$$\begin{aligned}de\,{}_3F_2\left(\begin{matrix}c & d & a-1\\ & d & e\end{matrix};\ 1\right) &= c(d+e-a-b-c){}_3F_2\left(\begin{matrix}c+1 & b+1 & a\\ & d+1 & e+1\end{matrix};\ 1\right) \\ &\quad + (d-c)(e-c){}_3F_2\left(\begin{matrix}c & b+1 & a\\ d+1 & e+1\end{matrix};\ 1\right).\end{aligned}$$

Finally equation 7 in [21] gives

$$\begin{aligned}{}_3F_2\left(\begin{matrix}c & b+1 & a\\ d+1 & e+1\end{matrix};\ 1\right) &= \frac{e}{e-c}{}_3F_2\left(\begin{matrix}c & b+1 & a\\ & d+1 & e\end{matrix};\ 1\right) \\ &\quad - \frac{c}{e-c}{}_3F_2\left(\begin{matrix}c+1 & b+1 & a\\ & d+1 & e+1\end{matrix};\ 1\right).\end{aligned}$$

Utilizing the above formulas and simplifying yields the result. ∎

LEMMA 3.3. *Let $r_i^{n,k}$ and $u_{n,m}^k$ be as above. Then for $n-2m \geq 2k+1$, $k = 0, 1,$ or 2,*

$$\begin{aligned}\langle r_k^{n,k}, u_{n,2m}^k\rangle &= \frac{(-1)^{m+k}2^{2k+2-n}(n-2k-1)!(2k+1)!(n-2m)!(2m+2k)!}{(n+2k+1)!(n-k-m)!(m+k)!} \\ &\quad \times {}_3F_2\left(\begin{matrix}-k & \frac{n-2m+2}{2} & \frac{n-2m+1}{2}\\ \frac{-2m-2k+1}{2} & n-m-k+1\end{matrix};\ 1\right).\end{aligned} \tag{3.11}$$

Proof. Although this result is true more generally we will use it only for the cases indicated, which simplifies the proof considerably. The above formula can be verified by hand for $m = 0,\ k \in \{0, 1, 2\}$ and $n \in \{2k + 1, 2k + 2\}$. With m even set $i = k$ in (3.10) to obtain

$$\langle r_k^{n+1,k}, u_{n+1,m}^k\rangle = \langle r_k^{n,k}, u_{n,m}^k\rangle - \epsilon_{k,n+1,m}\langle r_k^k, \phi_{n+1}^k\rangle. \tag{3.12}$$

From (3.1) and (2.10) we find that

$$\epsilon_{k,n+1,m}\langle r_k^k, \phi_n^k\rangle = \frac{(-1)^{\frac{m}{2}+k+2}2^{2k+1-n}(n-2k-1)!(2k-3)!(n-m)!(m+2k+2)!}{(n+2k+2)!(\frac{2n-2k-m}{2})!(\frac{m+2k+2}{2})!}$$
$$\times\, {}_3F_2\left(\begin{matrix}-k-1 & \frac{n-m+2}{2} & \frac{n-m+1}{2}\\ \frac{-m-2k-1}{2} & \frac{2n-m-2k+2}{2}\end{matrix};\ 1\right).$$

Set $a = \frac{n-m+2}{2}$, $b = -k-1$, $c = a - \frac{1}{2}$, and $d = \frac{-m-2k-1}{2}$. Then the hypergeometric functions associated with $\langle r_k^{n,k}, u_{n,m}^k\rangle$, $\langle r_k^{n+1,k}, u_{n+1,m}^k\rangle$, and $\epsilon_{k,n+1,m}$ are ${}_3F_2\left(\begin{matrix}a & b+1 & c\\ d+1 & e\end{matrix};\ 1\right)$, ${}_3F_2\left(\begin{matrix}a & b+1 & c+1\\ d+1 & e+1\end{matrix};\ 1\right)$, and ${}_3F_2\left(\begin{matrix}a & b & c\\ d & e\end{matrix};\ 1\right)$, respectively. From Lemma 3.2 we find that (3.11) satisfies (3.12) for $n \geq 2k + 1 + 2m$. Multiply (3.7) by r_k^k and integrate to obtain

$$\langle r_k^{n,k}, u_{n,2m}^k\rangle = \langle r_k^{n-2,k}, u_{n-2,n-2m-2}^k\rangle - \frac{c_{k,n-1,n-2m}}{\langle \phi_{n-1}^k, \phi_{n-1}^k\rangle}\langle r_k^k, \phi_{n-1}^k\rangle. \tag{3.13}$$

Since

$$c_{k,n-1,n-2m}\frac{\langle r_k^k, \phi_{n-1}^k\rangle}{\langle \phi_{n-1}^k, \phi_{n-1}^k\rangle} = \frac{(-1)^{m+k}2^{2k-n+3}(2n-1)(2k+1)!(n-2k-3)!(2k+2m)!}{(n-2k+1)!(n-k-m-1)!(m+k)!}$$
$$\times\, {}_3F_2\left(\begin{matrix}-k-1 & \frac{n-2m+2}{2} & \frac{n-2m+1}{2}\\ \frac{-2m-2k+1}{2} & n-m-k\end{matrix};\ 1\right), \tag{3.14}$$

it is not difficult to show that (3.11) satisfies (3.13). To see this interchange d and e in Lemma 3.2, eliminate ${}_3F_2\left(\begin{matrix}a & b+1 & c+1\\ d+1 & c+1\end{matrix};\ 1\right)$ and make the substitutions $a = \frac{n-2m+2}{2}$, $b = -k-1$, $c = \frac{n-2m+1}{2}$, $d = n-m-k$, and $e = \frac{-2m-2k+1}{2}$. Now multiply by

$$\frac{(-1)^{m+k}2^{2k-n+3}(2n-1)(2k+1)!(n-2k-3)!(2k+2m)!}{(n-2k+1)!(n-k-m-1)!(m+k)!}$$

and use (3.11) and (3.14). The result follows since (3.11) satisfies (3.12), (3.13), and the initial conditions mentioned above. □

The relation

$$\langle r_i^{n,k}, u_{n,2m+1}^k\rangle = \langle r_i^{n+1,k}, u_{n+1,2m+2}^k\rangle \tag{3.15}$$

shows that we need only compute the above inner products for m even. This relation can be obtained by observing that $\bar{u}_{n-(2m+1)}^k = \bar{u}_{(n+1)-(2m+2)}^k$ and that from the parity of $\bar{u}_{n+1-2m-2}^k$, $\langle \bar{u}_{n+1-2m-2}^k, \phi_{n+1}^k\rangle = 0$. Since $l_i^k(t) = r_i^k(-t)$ the parity of $u_{n,m}^k$ implies that

$$\langle l_i^{n,k}, u_{n,m}^k\rangle = (-1)^{n+m+1}\langle r_i^{n,k}, u_{n,m}^k\rangle. \tag{3.16}$$

We finish this section by obtaining a recurrence formula for the inner products $\langle u^k_{n,j}, u^k_{n,i}\rangle$. We will do this only for i and j even since the same reasoning as above shows $\langle u^k_{n,2j+1}, u^k_{n,2i+1}\rangle = \langle u^k_{n+1,2j+2}, u^k_{n+1,2i+2}\rangle$. Set $m = 2i$ and increment n by one in (3.3). Then multiply by $u^k_{n,2j}$ and integrate by parts to get

$$\begin{aligned}&-\langle u^k_{n+1,2i}(t), D[(1-t^2)u^k_{n,2j}(t)]\rangle\\ &\quad = -(n-2i)\langle u^k_{n+2,2i}, u^k_{n,2j}\rangle + (n-2i)\langle u^k_{n,2i}, u^k_{n,2j}\rangle + \delta_{k,n+1,2i}c_{k,n+1,n-2j}.\end{aligned}$$

From (3.5) we find that the term on the left-hand side of the above equation can be eliminated, giving

$$\begin{aligned}&-(n-2j)\langle u^k_{n+1,2i}, u^k_{n-1,2j}\rangle + (n-2j+2)\langle u^k_{n+1,2i}, u^k_{n+1,2j}\rangle\\ &= -(n-2i)\langle u^k_{n+2,2i}, u^k_{n,2j}\rangle + (n-2i)\langle u^k_{n,2i}, u^k_{n,2j}\rangle + \delta_{k,n+1,2i}c_{k,n+1,n-2j}.\end{aligned}$$

Another useful equation is

$$\langle u^k_{n+2,2i}, u^k_{n,2j}\rangle = \langle u^k_{n+1,2i}, u^k_{n+1,2j}\rangle + \epsilon_{k,n+2,2i}c_{k,n+1,n-2j},$$

which can be obtained from (3.1). The above two equations can be combined to give

$$(2n-2j-2i+3)\langle u^k_{n+1,2i}, u^k_{n+1,2j}\rangle = (2n-2j-2i+1)\langle u^k_{n,2i}, u^k_{n,2j}\rangle + \kappa_{k,n,2i,2j}, \tag{3.17}$$

where

$$\kappa_{k,n,2i,2j} = (\delta_{k,n+1,2i} - (n-2i+1)\epsilon_{k,n+2,2i})c_{k,n+1,n-2j} + (n-2j)\epsilon_{k,n+1,2i}c_{k,n,n-1-2j}. \tag{3.18}$$

4. Wavelets. Before we begin the actual construction of spline wavelets some general results will be given that will help in the construction and also show how to modify the bases constructed to obtain wavelet bases for compact intervals. Let the multiresolution analysis (V_p) be generated by n orthonormal scaling functions $\phi^0, \ldots, \phi^{n-1}$ minimally supported on $[-1, 1]$, with exactly k of these functions, $\phi^0, \ldots, \phi^{k-1}$ not having support in $[0, 1]$. Then from the theory of paraunitary operators there are n orthonormal wavelet functions $\psi^0, \ldots, \psi^{n-1}$, which we may assume are also minimally supported on $[-1, 1]$. Below, we give a method for constructing the wavelet functions from the scaling functions. In the case of symmetric or antisymmetric scaling functions, this method allows the construction of symmetric or antisymmetric wavelet functions.

Using the notation given in (1.4), define $Q_0 = \operatorname{span}\{\phi^0, \ldots, \phi^{k-1}\}$, $Q_1 = \operatorname{span}\{\phi^0_{1,0}, \ldots, \phi^{k-1}_{1,0}\}$, $Y = Q_0 \cap Q_1$, and $\dim Y = m$. Also, let $Y_0 = Q_0\chi_{[-1,0]} \cap Q_1\chi_{[-1,0]}$ and $Y_1 = Q_0\chi_{[0,1]} \cap Q_1\chi_{[0,1]}$ with dimensions m_0 and m_1, respectively.

LEMMA 4.1. *The number of wavelet functions ψ^j such that $P_{Q_1}\psi^j \neq 0$ is at least $k - m$.*

Proof. Suppose that there are $\hat{k} < k - m$ such wavelet functions $\psi^0, \ldots, \psi^{\hat{k}-1}$. Suppose then that $v \in Q_1 \ominus Y$ is chosen to have unit length and to be orthogonal to $P_{Q_1}\operatorname{span}\{\psi^0, \ldots, \psi^{\hat{k}-1}\}$. Such a choice is possible because $Q_1 \ominus Y$ has $k - m$ dimensions while $P_{Q_1}\operatorname{span}\{\psi^0, \ldots, \psi^{\hat{k}-1}\}$ has at most $\hat{k}$ dimensions. Let $y = (I - P_{Q_0})v \in V_1 \ominus V_0$. Since $\operatorname{span}\{\phi^k, \ldots, \phi^{n-1}\} \perp Q_1$, (1.1) implies that $y \in W_0$.

On the other hand, we claim that y cannot be written as a linear combination of the wavelet functions. Indeed, suppose y can be expressed as a linear combination of

wavelet functions. Then in particular $P_{Q_1}y$ must be in $P_{Q_1}\text{span}\{\psi^0,\dots,\psi^{\hat{k}-1}\}$ and hence orthogonal to v. That is,

$$\begin{aligned}0 &= P_v P_{Q_1} y\\ &= P_v P_{Q_1}(I - P_{Q_0})v\\ &= P_v(P_{Q_1} - P_{Q_1}P_{Q_0})v\\ &= P_v(v - P_{Q_1}P_{Q_0}v)\\ &= v - P_v P_{Q_1}P_{Q_0}v\end{aligned}$$

or

$$v = P_v P_{Q_1} P_{Q_0} v.$$

Because these projections are orthogonal, the above equality is possible only if $v \in Q_0$, which is not the case since $v \in Q_1 \ominus Y$. □

LEMMA 4.2. *The number of wavelet functions not supported on* $[0,1]$ *is at least* $j = 2k - m_0 - m_1$.

Proof. Suppose that there are fewer than j such wavelet functions. By the above lemma, there must be at least $k-m$ of them, whose span will be denoted by Ψ^{s}, such that $P_{Q_1}\psi \neq 0$ for $\psi \in \Psi^{\text{s}} \setminus \{0\}$. In addition, there are $\hat{\jmath} < k - m_0 - m_1 + m$ others, spanning Ψ^{a}. We assume that Ψ^{s} has exactly $k-m$ dimensions and $P_{Q_1}\Psi^{\text{a}} = \{0\}$. Since every wavelet function is orthogonal to the generators of V_0,

$$P_{Q_1}(\Psi^{\text{s}} \oplus \Psi^{\text{a}}) = P_{Q_1 \ominus Y}(\Psi^{\text{s}} \oplus \Psi^{\text{a}}),$$

where $Q_1 \ominus Y$ has at most $k-m$ dimensions. Thus, by taking linear combinations if necessary, one can always arrange for this assumption to be correct. Let Ψ^1 be the span of the remaining wavelets supported on $[0,1]$.

We begin by defining the following five spaces:

$$\begin{aligned}T &= (I - P_{Q_1})Q_0,\\ T_0 &= \{t \in T\colon\ \text{supp}\, t \subset [-1,0]\},\\ T_1 &= \{t \in T\colon\ \text{supp}\, t \subset [0,1]\},\\ U &= T \ominus T_0 \ominus T_1,\\ Z &= (\chi_{[0,1]} - \chi_{[-1,0]})U\end{aligned}$$

for which the dimensions are $k-m$, m_0-m, m_1-m, $k-m_0-m_1+m$, and $k-m_0-m_1+m$, respectively. Note that since $U \perp Q_1$, multiplying by characteristic functions to construct Z does not destroy the membership of these functions in V_1. Also, since the supports of T_0 and T_1 are $[-1,0]$ and $[0,1]$, it follows that these spaces are orthogonal to U on each of said intervals and thus are orthogonal to Z.

Next, we observe that $Q_0 \oplus \Psi^{\text{s}} = T \oplus Q_1$. It is easy to show that both spaces have dimension $2k-m$. Furthermore, $T \oplus Q_1 = Q_0 + Q_1 \subset Q_0 \oplus \Psi^{\text{s}}$ because among the wavelet functions and their translates, only those in Ψ^{s} are not orthogonal to Q_1.

Now, by its construction, Z is orthogonal to any translates of Ψ^1 as well as nonzero translates of Ψ^{s}, Ψ^{a}, and Q_0. It is also orthogonal to arbitrary translates of the scaling functions supported on $[0,1]$. However, $Z \subset V_1$, so we have

$$\begin{aligned}Z &\subset Q_0 \oplus \Psi^{\text{s}} \oplus \Psi^{\text{a}}\\ &= T \oplus Q_1 \oplus \Psi^{\text{a}}\\ &= U \oplus T_0 \oplus T_1 \oplus Q_1 \oplus \Psi^{\text{a}},\end{aligned}$$

and since $Z \perp T_0 \oplus T_1 \oplus Q_1$, it follows that

$$Z \subset U \oplus \Psi^{\mathrm{a}}.$$

But Z has more dimensions than Ψ^{a}, so there exists a nonzero $z \in Z$ which is also in U. By the definition of Z, there is some $u \in U$ such that $(\chi_{[0,1]} - \chi_{[-1,0]})u = z$. Finally, consider the function $x = u + z$. Since both u and z are in U, we know that $x \in U \perp T_1$. On the other hand, $x \in U \subset T$, and clearly $\operatorname{supp} x \subset [0,1]$, so $x \in T_1$. The only way for both of these to be true is to have $x = 0$, which yields a contradiction. ∎

For the scaling vectors constructed in the next section, we find that $m_0 = m_1 = m = 0$. In this event we have the following corollary.

COROLLARY 4.3. *If $m_0 = m_1 = m = 0$, then the number of wavelet functions not supported on $[0,1]$ is at least $2k$.*

From the above lemmas we have the following theorem.

THEOREM 4.4. *Let the multiresolution analysis (V_p) be generated by n orthonormal scaling functions $\phi^0, \ldots, \phi^{n-1}$ minimally supported on $[-1,1]$. Let k, m, m_0, and m_1 be defined as above. Then there exists n orthonormal wavelet functions $\psi^0, \ldots, \psi^{n-1}$ minimally supported on $[-1,1]$ such that exactly $2k - m_0 - m_1$ are not supported in $[0,1]$.*

Proof. As indicated above, $k - m$ wavelets Ψ^{s} not supported in $[0,1]$ may be constructed as an orthonormal basis for $(I - P_{Q_0})Q_1$ since this basis is in V_1 and is orthogonal to V_0. We construct $k - m_0 - m_1 + m$ more functions in the following way. As indicated in Lemma 4.2, Z is orthogonal to $V_0 \ominus \big(Q_0 \ominus (Q_0 \cap Q_1)\big)$ as well as any wavelet functions supported on $[0,1]$. Thus the wavelet functions we seek are found by taking an orthonormal basis for $\operatorname{span}(I - P_{Q_0 \cup \Psi^{\mathrm{s}}})Z$.

We now show that the remaining $n - 2k + m_0 + m_1$ may be chosen so that they are supported in $[0,1]$. Since the dimensions of A_0 and A_1 are $n-k$ and $2n-k$, respectively, $A_1 \ominus A_0$ is n-dimensional. From this space we select a subspace K perpendicular to $\{\phi^i|_{[0,1]}\}_{i=0}^{k-1}$ and $\{\phi^i(\cdot - 1)|_{[0,1]}\}_{i=0}^{k-1}$. The dimension of this space will determine the maximum number of wavelets with support contained in $[0,1]$. If after an appropriate change of basis $\phi^0|_{[0,1]}$ is in Y_0, then ϕ^0 is perpendicular to $A_1 \ominus A_0$. Thus K needs to be chosen perpendicular only to a $(k-1)$-dimensional subspace of $\operatorname{span}\{\phi^i|_{[0,1]}\}_{i=1}^{k-1}$ and a k-dimensional subspace of $\operatorname{span}\{\phi^i(\cdot - 1)|_{[0,1]}\}_{i=0}^{k-1}$. Proceeding in this way we find that the dimension of K is $n - 2k + m_0 + m_1$, which completes the proof. ∎

Suppose Φ is a set of n scaling functions all supported on $[-1,1]$ and let Ψ be a set of wavelet functions constructed as above. Since the wavelet functions are all supported on $[-1,1]$ they satisfy the following equation:

$$\Psi(t) = \sum_{i=-2}^{1} D_i \Phi(2t - i), \tag{4.1}$$

where each D_i is an $n \times n$ matrix.

The above decomposition of the wavelet basis allows a description of how to obtain a basis when the multiresolution analysis is restricted to $[0,1]$. To this end let Ψ^{s}, Ψ^{a}, and Ψ^1 be as above and set $\Psi_{n,j} = \Psi(2^n \cdot - j)$.

THEOREM 4.5. *Let the multiresolution analysis (V_p) be generated by orthogonal scaling functions $\Phi = \{\phi^0, \ldots, \phi^{n-1}\}$, minimally supported on $[-1,1]$, and let k, m, m_0, and m_1 be defined as above. If $\Phi\chi_{[0,1]}$ is an orthogonal set, then for $q \geq 0$ $\{\phi^i_{q,j}\chi_{[0,1]}\colon\ i = 0, \ldots, n-1,\ j = 0, \ldots, 2^q\} \backslash \{0\}$ is an orthogonal basis for $\bar{V}_q = V_q\chi_{[0,1]}$.*

Furthermore, there exist orthogonal bases

$$\bar{\Psi}_q = \left(\bigcup_{j=0}^{2^q-1} \Psi^1_{q,j} \cup \bigcup_{j=0}^{2^q} \Psi^2_{q,j} \cup \bigcup_{j=0}^{2^q-1} \Psi^3_{q,j} \cup \bigcup_{j=1}^{2^q} \Psi^4_{q,j} \cup \bigcup_{j=1}^{2^q-1} \Psi^5_{q,j} \right) \chi_{[0,1]} \tag{4.2}$$

for the wavelet spaces $\bar{W}_q = W_q\chi_{[0,1]}$ *such that* $\bar{V}_0 \oplus \bigoplus_{q\geq 0} \bar{W}_q = L^2[0,1]$.

Proof. First note that if $\Phi\chi_{[0,1]}$ is orthogonal, then $\Phi\chi_{[-1,0]}$ is as well. From this it is easy to see that $\{\phi^i_{q,j}\chi_{[0,1]}\colon\ i = 0,\dots,n-1,\ j = 0,\dots,2^q\} \setminus \{0\}$ is an orthogonal basis for $\bar{V}_q$.

Now, for the wavelet spaces, we know that $\bar{W}_q = \bar{V}_{q+1} \ominus \bar{V}_q$. Thus, if $\bar{\Psi}$ is any orthogonal basis for $\Psi^{\mathrm{a}} \oplus \Psi^{\mathrm{s}}$, a basis for $\bar{W}_q$ may include

$$\bigcup_{j=0}^{2^q-1} \Psi^1_{q,j} \cup \bigcup_{j=1}^{2^q-1} \bar{\Psi}_{q,j}.$$

To complete the basis, however, it is necessary to see what happens near the ends of the interval $[0,1]$ and to choose a suitable $\bar{\Psi}$.

Consider first the left endpoint. Since the remaining basis functions must be orthogonal to those given above as well as the generators for $\bar{V}_q$, it is clear that they must be obtained from span $\bar{\Psi}_{q,0}$ by truncation to the interval $[0,1]$. In addition, they must be orthogonal to $\Phi_{q,0}\chi_{[0,1]}$. The number of functions needed can be found by counting the dimensions associated with their support and subtracting the number of orthogonality restrictions. For the former, note that $q \geq 0$ $\bar{\Psi}_{q,0}\chi_{[0,1]} \subset \operatorname{span}(\Phi_{q+1,0}\chi_{[0,1]} \cup \Phi_{q+1,1})$, a $2n$-dimensional space. For the latter, note that these functions must be orthogonal to $\operatorname{span}(\Phi_{q,0}\chi_{[0,1]} \cup \Psi^1_{q,0})$, which has $2n-2k+m_0+m_1$ dimensions, as well as a $(k-m_0)$-dimensional subspace of span $\Phi_{q,1}$. Thus $2n-(2n-2k+m_0+m_1)-(k-m_0) = k-m_1$ wavelet functions are required to complete the basis on the left. Similar arguments show that there are $k-m_0$ on the right and that these two groups have $\hat{k} = \max\{k-m_0-m_1, 0\}$ functions in common (before truncation). We remark that while it is unusual for an MRA to have nonzero m, m_0, or m_1, it is even more unusual, but still possible, to have $k-m_0-m_1 < 0$.

Finally, we choose $\bar{\Psi}$ so that it can be partitioned into four sets:

- Ψ^2, with $\hat{k}$ functions used on both ends,
- Ψ^3, with $k-m_1-\hat{k}$ functions used only on the left,
- Ψ^4, with $k-m_0-\hat{k}$ functions used only on the right, and
- Ψ^5, with $\hat{k}$ functions not used on either end.

With Ψ^2, Ψ^3, Ψ^4, and Ψ^5 defined in this fashion, (4.2) gives an orthogonal basis for $\bar{W}_q$. ∎

5. Construction of continuous spline wavelets. We now set $k = 0$ so that S^n_0 is the space of piecewise C^0 polynomials with knots at the integers and $V^{n,0}_0 = S^n_0 \cap L^2(\mathbf{R})$. By Lemma 2.1, $\{\phi^0_2,\dots,\phi^0_n\}$ forms an orthogonal basis for $A^{n,0}_0(\frac{\cdot+1}{2})$, where $\phi^0_n(t) = (1-t^2)p^{5/2}_{n-2}(t)$. In this case $r^0_0(t) = 1+t$ while $l^0_0(t) = 1-t$, and $C^{n,0}_0(\frac{\cdot+1}{2})$ and $C^{n,0}_1(\frac{\cdot+1}{2})$ are each one dimensional. Hence from Theorem 1.1 we see that an orthogonal intertwining MRA can be constructed if we can find a function $w \in A^{n,0}_1(\frac{\cdot+1}{2}) \ominus A^{n,0}_0(\frac{\cdot+1}{2})$ satisfying $\langle (I-P_w)r^{n,0}_0, (I-P_w)l^{n,0}_0\rangle = 0$, which for nonzero w is equivalent to

$$\langle r^{n,0}_0, l^{n,0}_0\rangle\langle w,w\rangle = \langle r^{n,0}_0, w\rangle\langle w, l^{n,0}_0\rangle. \tag{5.1}$$

As a member of $A_0^{n,0}(\frac{\cdot+1}{2})$, w must be a spline of degree n with knots at the integers. Since we would like to construct wavelets that are symmetric or antisymmetric we shall choose w so that it is symmetric or antisymmetric and also so that the knot at zero is as smooth as possible. Since the dimension of the space of piecewise polynomial functions in $A_1^{n,0}(\frac{\cdot+1}{2}) \ominus A_0^{n,0}(\frac{\cdot+1}{2})$ that are C^{n-2} at zero is two, there exists a basis x_1, x_2 for this space where x_1 is symmetric and x_2 is antisymmetric (one of them being a multiple of $u_{n,0}$). Thus if we wish to find a symmetric or antisymmetric w satisfying (5.1), it will be at most C^{n-3} at zero. Naturally, if symmetry is not important, a smoother w may be constructed from the space spanned by x_1 and x_2. Because of (5.3) below we see that w must be chosen having the same parity as $(-1)^{n+1}$. Thus for symmetry and the greatest possible smoothness we are forced to choose

$$w_n^0 = \alpha_0(n) u_{n,0}^0 + u_{n,2}^0 \tag{5.2}$$

for $n \geq 3$. From (3.8), (3.11), and (3.16) we have

$$\langle r_0^{n,0}, l_0^{n,0} \rangle = \frac{(-1)^{n+1} 8}{(n+2)(n+1)n}, \tag{5.3}$$

$$\langle r_0^{n,0}, u_{n,2m}^0 \rangle = \frac{(-1)^m (n-1)!(n-2m)!(2m)!}{2^{n-2}(n+1)!(n-m)!(m)!}, \tag{5.4}$$

and

$$\langle l_0^{n,0}, u_{n,m}^0 \rangle = (-1)^{m+n+1} \langle r_0^{n,0}, u_{n,m}^0 \rangle. \tag{5.5}$$

Also from (3.9) we find

$$\langle r_0^{n,0}, r_0^{n,0} \rangle = (n+1) \frac{\langle r_0^0, \phi_n^0 \rangle \langle r_0^0, \phi_{n+1}^0 \rangle}{\langle \phi_n^0, \phi_n^0 \rangle} = \frac{8}{n(n+2)}. \tag{5.6}$$

With this (5.1) becomes

$$\begin{aligned} 0 = &\begin{vmatrix} \langle u_{n,0}^0, u_{n,0}^0 \rangle & \langle r_0^{n,0}, u_{n,0}^0 \rangle \\ \langle r_0^{n,0}, u_{n,0}^0 \rangle & |\langle r_0^{n,0}, l_0^{n,0} \rangle| \end{vmatrix} \alpha_0(n)^2 + 2 \begin{vmatrix} \langle u_{n,0}^0, u_{n,2}^0 \rangle & \langle r_0^{n,0}, u_{n,2}^0 \rangle \\ \langle r_0^{n,0}, u_{n,0}^0 \rangle & |\langle r_0^{n,0}, l_0^{n,0} \rangle| \end{vmatrix} \alpha_0(n) \\ &+ \begin{vmatrix} \langle u_{n,2}^0, u_{n,2}^0 \rangle & \langle r_0^{n,0}, u_{n,2}^0 \rangle \\ \langle r_0^{n,0}, u_{n,2}^0 \rangle & |\langle r_0^{n,0}, l_0^{n,0} \rangle| \end{vmatrix}, \end{aligned} \tag{5.7}$$

where we have used the sign structure of (5.3) and (5.5). In order to find solutions to this equation we need the following formula, for $n \geq 2\max\{i,j\}+1$ with $i,j \in \{0,2\}$:

$$\langle u_{n,2i}^0, u_{n,2j}^0 \rangle = \frac{(-1)^{i+j}(n-2i)!(n-2j)!(2i)!(2j)!(n-1)!(n^2+5n+2-2i-2j)}{2^{2n-1} i!j!(n-i)!(n-j)!(n+1)!(2n+1-2i-2j)}. \tag{5.8}$$

For the cases $n > 2\max\{i,j\}+1$ the above formula follows by induction using (3.17) with

$$\kappa_{0,n,2i,2j} = \frac{(-1)^{i+j+1}(2i)!(2j)!(n-2i)!(n-2j)!(n-1)!}{2(2n-1)i!j!(n+1-i)!(n+1-j)!(n+2)!} \Pi_n,$$

where

$$\begin{aligned} \Pi_n = &3n^5 + (27-i-j)n^4 + (85-14(i+j))n^3 + (117-51(i+j)+2(i+j)^2)n^2 \\ &+ (72-62(i+j)+10(i^2+j^2)+24ij)n \\ &+ 16 - 24(i+j) - 4ij(i+j) + 20ij + 8(i^2+j^2), \end{aligned}$$

as well as the initial conditions $\langle u^0_{2,0}, u^0_{2,0}\rangle = \frac{1}{15}$, $\langle u^0_{4,2}, u^0_{4,0}\rangle = -\frac{17}{13440}$, and $\langle u^0_{4,2}, u^0_{4,2}\rangle = \frac{1}{960}$. For $n = 3, i = 0, j = 1$ and $n = 3, i = 1, j = 1$ (5.8) can be verified by hand.

Substituting the above formulas into (5.7) and solving for $\alpha_0(n)$ yields two solutions,

$$\alpha_0(n) = 2\frac{\frac{(n-2)(2n+1)}{(n-1)} \pm 2\sqrt{3}\sqrt{\frac{(2n+1)(n+1)}{(2n-3)(n-1)}}}{n(2n-1)}, \tag{5.9}$$

either of which will suffice to give w as in (5.2). This w is then used to construct orthogonal the scaling functions. For $j = 2, \ldots, n$, set

$$\tilde{\phi}^j = \begin{cases} \phi^0_j(2\cdot - 1) & \text{if } t \in [0,1), \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{\phi}^1 = \begin{cases} w(2\cdot - 1) & \text{if } t \in [0,1), \\ 0 & \text{otherwise,} \end{cases}$$

and $\tilde{\phi}^0 = (I - P_{\{\tilde{\phi}^1, \ldots, \tilde{\phi}^n, \tilde{\phi}^1(\cdot+1), \ldots, \tilde{\phi}^n(\cdot+1)\}})h$, where

$$h(t) = \begin{cases} (1 - |t|) & \text{if } t \in [-1,1), \\ 0 & \text{otherwise,} \end{cases}$$

where P_x is the orthogonal projection onto the subspace spanned by x.

THEOREM 5.1 (see [7], [8]). *For $n \geq 3$ and $\alpha_0(n)$ given by* (5.9), $\tilde{\Phi} = \{\tilde{\phi}^0, \ldots, \tilde{\phi}^n\}^\top$ *generates an orthogonal multiresolution analysis $\{\tilde{V}^{n,0}_k\}$. Furthermore, the last n functions are symmetric or antisymmetric about $\frac{1}{2}$ while the first function is symmetric about* 0.

Proof. Since $V^{n,0}_0 = \mathrm{cl}_{L^2}\,\mathrm{span}\{h(\cdot - i), \tilde{\phi}^2(\cdot - i), \ldots, \tilde{\phi}^n(\cdot - i)\ \forall i \in \mathbf{Z}\}$ and $\tilde{V}^{n,0}_0 = \mathrm{cl}_{L^2}\,\mathrm{span}\{\tilde{\phi}^0(\cdot - i), \ldots, \tilde{\phi}^n(\cdot - i)\ \forall i \in \mathbf{Z}\}$ we find that $V^{n,0}_0 \subset \tilde{V}^{n,0}_0$. The result now follows from the above construction.

Remark. As we have shown, the generators of $\tilde{\Phi}$ are the smoothest functions derived from Φ having the support, symmetry, and orthogonality properties indicated above.

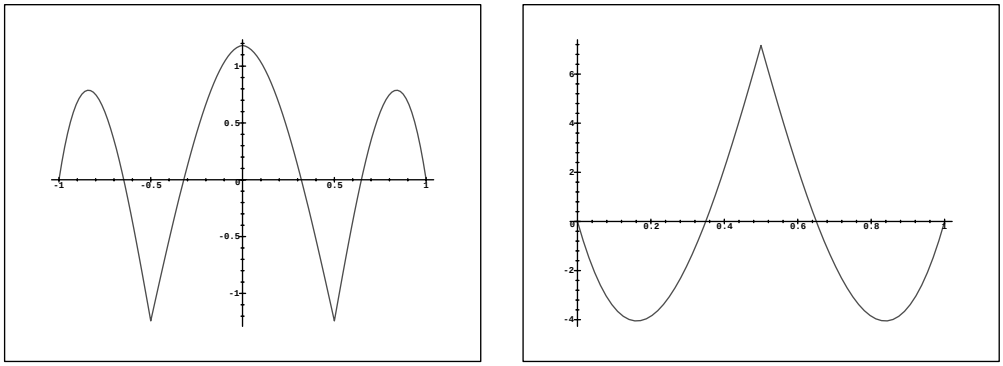
With the scaling functions above we may now construct the coefficients $C^0_{n,i}$, $i = -2, -1, 0, 1$ in the matrix refinement equation. In light of Theorem 4.4, two of the wavelets will not be supported only in $[0,1]$ while the remaining $n - 1$ will be supported on $[0,1]$. In particular, the following holds.

COROLLARY 5.2. *Let*

$$\tilde{\psi}^0 = \sqrt{2}(I - P_{\tilde{\phi}^0})\tilde{\phi}^0_{1,0},$$
$$\tilde{\psi}^1 = \sqrt{2}(\chi_{[0,1]} - \chi_{[-1,0]})(I - P_{\tilde{\phi}^0_{1,0}})\tilde{\phi}^0,$$

and $\{\tilde{\psi}^i\colon\ i = 2, \ldots, n\}$ be an orthonormal basis for Ψ^1 consisting of functions that are symmetric or antisymmetric with respect to $\frac{1}{2}$. Then $\{\tilde{\psi}^0, \ldots, \tilde{\psi}^n\}$ generates a shift-invariant orthonormal basis for W_0. Furthermore, $\tilde{\psi}^0(0) = \tilde{\phi}^0(0)$.

Proof. Up to the constants, the formulas for $\tilde{\psi}^0$ and $\tilde{\psi}^1$ follow directly from Theorem 4.4. It is true in general that $\langle \tilde{\phi}^0, \tilde{\phi}^0_{0,1}\rangle = \frac{1}{\sqrt{2}}$, and hence the normalization factors for both $\tilde{\psi}^0$ and $\tilde{\psi}^1$ are as shown. To see why this is the case, observe that $\tilde{\phi}^0_{1,0}$ is the only generator of V_1 that does not vanish at 0, and $\tilde{\phi}^0_{1,0}(0) = \sqrt{2}\phi^0(0)$. It is also for this reason that $\tilde{\psi}^0(0) = \tilde{\phi}^0(0)$.



$\tilde{\phi}^0$ $\tilde{\phi}^1$

FIG. 1. *Selected C_0 scaling functions with approximation order 4 ($n = 3$).*

TABLE 1
Formulas for the scaling functions.

$$\tilde{\phi}^0(t) = \begin{cases} \frac{\sqrt{30}\sqrt{327+56\sqrt{14}}(1855-256\sqrt{14})}{25235210}(2002t^3 - (645\sqrt{14}+658)t^2 & \\ \quad -(1106-285\sqrt{14})t + 322 - 16\sqrt{14}) & \text{for } 0 \le t < \frac{1}{2} \\ \frac{\sqrt{30}\sqrt{327+56\sqrt{14}}(-6433+1040\sqrt{14})}{25235210}(20818t^2 + (1835\sqrt{14}-17024)t & \\ \quad +2254 - 1176\sqrt{14})(t-1) & \\ & \text{for } \frac{1}{2} \le t \le 1 \\ \tilde{\phi}^0(-t) & \text{for } -1 \le t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^1(t) = \begin{cases} \frac{3\sqrt{15}(\sqrt{7}+21\sqrt{2})}{1750}(280t^2 - (75\sqrt{14}+665)t + 322 + 54\sqrt{14})t & \text{for } 0 \le t \le \frac{1}{2} \\ \tilde{\phi}^1(1-t) & \text{for } \frac{1}{2} < t \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^2(t) = \sqrt{30}t(1-t)\chi_{[0,1]}$$

$$\tilde{\phi}^3(t) = \sqrt{210}t(1-t)(2t-1)\chi_{[0,1]}$$

In the examples given by Theorem 5.1, Ψ^1 is symmetrical about $\frac{1}{2}$, and hence it must have an orthonormal basis consisting of symmetrical and antisymmetrical functions. □

An example is given in Figure 1, where the minus sign is chosen in (5.9), and the analytic formulas associated with this example can be found in Table 1.[1] The wavelets with support $[-1, 1]$ may be computed using Corollary 5.2 while those supported in $[0, 1]$ can be obtained by finding $n-1$ orthogonal functions, symmetrical or antisymmetrical with respect to $\frac{1}{2}$, from the space $(I - P_{\{\tilde{\phi}^0, \tilde{\phi}^0(\cdot-1)\}})A_1^{n,0} \ominus A_0^{n,0}$. The matrix coefficients in the refinement equation for the scaling function may be easily calculated using the orthogonality of these functions. This also holds for the matrix coefficients in the expansion for the wavelets.

As a final remark to this section we note that because of the symmetry of the scaling functions and wavelets, these bases can easily be modified to bases for compact intervals. Using the notation $\tilde{\phi}^i_{k,j}(x) = \tilde{\phi}^i(2^k x - j)$, let $\bar{\phi}^i_{k,j} = \tilde{\phi}^i_{k,j}|_{[0,1]}$, $\bar{\psi}^m_{k,j} = \tilde{\psi}^m_{k,j}|_{[0,1]}$

[1] Wavelets and the matrices in the refinement equation for this and other examples can be found at www.math.gatech.edu/∼geronimo.

for $m \neq 1$, and

$$\bar{\psi}^1_{k,j} = \begin{cases} 0 & \text{if } \operatorname{supp} \tilde{\psi}^1_{k,j} \cap [0,1]^c \neq \emptyset, \\ \tilde{\psi}^1_{k,j} & \text{otherwise;} \end{cases}$$

we find the following theorem.

THEOREM 5.3. *The set* $\{\bar{\phi}^i_{k,j}\colon\ k \geq 0,\ i = 1, \ldots, n,\ 0 \leq j \leq 2^k - 1 + \delta_{0,i}\}$ *is an orthogonal basis for* $\bar{V}^{n,0}_k = \tilde{V}^{n,0}_k \cap L^2[0,1]$ *while* $\{\bar{\psi}^i_{k,j}\colon\ k \geq 0,\ i = 1, \ldots, n,\ \delta_{1,i} \leq j \leq 2^k - 1 + \delta_{0,i}\}$ *forms an orthogonal basis for* $\bar{W}^{n,0}_k = \tilde{W}^{n,0}_k \cap L^2[0,1]$. *Furthermore,* $\operatorname{cl}_{L^2} \bar{V}^{n,0}_0 \oplus \bigoplus_{k \geq 0} \bar{W}^{n,0}_k = L^2[0,1]$.

6. Construction of differentiable spline wavelets. For differentiable splines, we take $k = 1$ and consider the space $V^{n,0}_1 = S^n_1 \cap L^2(\mathbf{R})$. Lemma 2.1 says that $\{\phi^1_4, \ldots, \phi^1_n\}$ forms an orthogonal basis for $A^{n,1}_0(\frac{\cdot+1}{2})$, where $\phi^1_n(t) = (1-t^2)^2 p^{9/2}_{n-4}(t)$. In this case we have $r^1_i(t) = (1+t)^2(1-t)^i$ and $l^1_i(t) = (1-t)^2(1+t)^i$, $i = 0, 1$. The spaces $C^{n,1}_0(\frac{\cdot+1}{2})$ and $C^{n,1}_1(\frac{\cdot+1}{2})$ are each two dimensional. Hence from Theorem 1.1 we see that an orthogonal intertwining MRA can be constructed if we can find two functions $w_1, w_2 \in A^{n,1}_1(\frac{\cdot+1}{2}) \ominus A^{n,1}_0(\frac{\cdot+1}{2})$ such that

$$\langle (I - P_{\{w_1,w_2\}}) r^{n,1}_i, (I - P_{\{w_1,w_2\}}) l^{n,1}_j \rangle = 0,\ i \leq j = 0, 1. \tag{6.1}$$

In order to construct scaling functions with a symmetry axis one of these functions will be constructed symmetric and the other antisymmetric. From (3.8) we find

$$\langle r^{n,1}_0, l^{n,1}_0 \rangle = \frac{128(-1)^{n+1}(n^2+2n-9)(n-2)!}{(n+3)!}, \tag{6.2}$$

$$\langle r^{n,1}_0, l^{n,1}_1 \rangle = \frac{768(-1)^{n+1}(n-2)!}{(n+3)!},$$

and

$$\langle r^{n,1}_1, l^{n,1}_1 \rangle = \frac{4608(-1)^{n+1}(n-3)!}{(n+4)!}. \tag{6.3}$$

With $k = 1$, (3.11) yields

$$\begin{aligned} \langle r^{n,1}_1, u^1_{n,2m} \rangle &= \langle r_{n,1}, u^1_{n,2m} \rangle \\ &= \frac{3(-1)^{m+1}(2m)!(n-2m)!(n^2+5n+2-8m)(n-3)!}{2^{(n-5)} m!(n-m)!(n+3)!}. \end{aligned} \tag{6.4}$$

Combining (6.4) with (3.7) and (3.10) and using initial condition $\langle r^{2,1}_0, u^1_{2,2} \rangle = \frac{10}{3}$, we obtain

$$\langle r^{n,1}_0, u^1_{n,2m} \rangle = \frac{(-1)^{m+1}(2m)!(n-2m)!(n^2+7n+4-12m)(n-2)!}{2^{(n-4)} m!(n-m)!(n+2)!}. \tag{6.5}$$

In order to simplify the computations somewhat we biorthogonalize the ramp functions. Set $r_{n,0} = r^{n,1}_0$, $l_{n,0} = l^{n,1}_0$, $r_{n,1} = r^{n,1}_1 - \frac{\langle r^{n,1}_1, l_{n,0} \rangle}{\langle r_{n,0}, l_{n,0} \rangle} r_{n,0}$, and $l_{n,1} = l^{n,1}_1 - \frac{\langle l^{n,1}_1, r_{n,0} \rangle}{\langle r_{n,01}, l_{n,1} \rangle} l_{n,1}$. With the help of the inner products given above, we find

$$\langle r_{n,1}, l_{n,1} \rangle = \frac{(-1)^n 4608(n-3)!}{(n+4)!(n^2+2n-9)}, \tag{6.6}$$

$$\langle r_{n,1}, u^1_{n,2m}\rangle = \frac{96(-1)^m(2m)!(n-2m)!(n-3)!(n^2-4mn+3n+6)}{2^n m!(n-m)!(n^2+2n-9)(n+3)!}, \tag{6.7}$$

and

$$\langle r_{n,1}, u^1_{n,2m+1}\rangle = \frac{96(-1)^m(2m+1)!(n-2m-1)!(n+3)(n-2)!(3n^2-4mn+9n-8m)}{2^n m!(n-m)!(n^2+2n-9)(n+4)!}. \tag{6.8}$$

As in the C^0 case we can use (3.17) to compute the inner products $\langle u^1_{n,2i}, u^1_{n,2j}\rangle$; $i, j = 0, 1$, or 2; and $n > 2\max\{i, j\} + 3$. This computation was done using Maple and yielded

$$\langle u^1_{n,2i}, u^1_{n,2j}\rangle = \frac{(-1)^{i+j}(n-2i)!(n-2j)!(2i)!(2j)!(n-3)!q(n,i,j)}{2^{2n-1}(2n+1-i-j)(n+3)!i!j!(n-i)!(n-j)!}, \tag{6.9}$$

where

$$\begin{aligned} q(n,i,j) = {} & n^6 + 19n^5 + (131 - 8(i+j))n^4 + (368 - 112(i+j))n^3 \\ & + (372 - 424(i+j) + 24(i+j)^2)n^2 \\ & + (212 - 320(i+j) + 120(i^2+j^2) + 432ij)n \\ & + 2(24 - 48((i+j)(1+ij) + (i^2+j^2)) + 81ij). \end{aligned}$$

We now construct three orthogonal functions $v_0 = b_{0,0}(n)u^1_{n,0} + b_{0,2}(n)u^1_{n,2} + u^1_{n,4}$, $v_2 = b_{2,0}(n)u^1_{n,0} + b_{2,2}(n)u^1_{n,2} + u^1_{n,4}$, and $v_4 = b_{4,0}(n)u^1_{n,0} + b_{4,2}(n)u^1_{n,2} + u^1_{n,4}$, with the additional constraints that v_0 be orthogonal to $r_{n,0}$ and $r_{n,1}$, v_2 be orthogonal to $r_{n,1}$ and v_0, and v_4 be orthogonal to v_0 and v_2. Using the inner product formulas and a symbolic manipulation package such as Maple, we find $b_{0,0}(n) = \frac{12}{(n-3)(n-2)}$, $b_{0,2}(n) = \frac{12(n-1)}{(n-3)(n-2)}$, $b_{2,0}(n) = \frac{12(2n+1)(7n^2-17n+18)}{(2n-7)(n-2)(n-3)(7n^2-n+18)}$, $b_{2,2}(n) = \frac{12(14n^2-25n+54)(n-1)^2}{(2n-7)(n-2)(n-3)(7n^2-n+18)}$,

$$b_{4,0}(n) = \frac{4(2n+1)(2n-1)(9n^5+175n^4+285n^3-5695n^2+12786n-8280)}{(2n-7)(2n-5)(n-3)(n-2)(n^2+11n-6)(3n^3+28n^2-67n+28)},$$

and

$$b_{4,2}(n) = \frac{4(n-1)(2n-1)(9n^5+179n^4+477n^3-4163n^2+7458n-4392)}{(2n-7)(n-3)(n-2)(n^2+11n-6)(3n^3+28n^2-67n+28)}.$$

These equations allow us to compute the following inner products between the new functions with the biorthogonal ramps:

$$\langle v_2, r_{n,0}\rangle = \frac{18432(n^2+2n-9)(n-4)!}{2^n(n+1)!(2n-7)(7n^2-n+18)},$$

$$\langle v_4, r_{n,0}\rangle = \frac{-7680(n^2+15n-24)}{2^n(2n-7)(2n-5)(n+2)n(n-3)(3n^3+28n^2-67n+28)},$$

and

$$\langle v_4, r_{n,1}\rangle = \frac{3072q(n)(n-1)(n-4)!}{2^n(2n-7)(2n-5)(n-2)(n+3)!(n^2+11n-6)(n^2+2n-9)(3n^3+28n^2-67n+28)},$$

where

$$q(n) = 107n^7 + 1230n^6 - 3580n^5 - 9546n^4 + 21437n^3 + 30204n^2 - 70956n + 25920.$$

Also, we have

$$\langle v_0(n), v_0(n)\rangle = \frac{11052(2n-9)!!}{2^{2n}(2n+1)!!(n-2)^2(n-3)^2},$$

$$\langle v_2(n), v_2(n)\rangle = \tfrac{11052(2n-9)!!q(n)}{2^{2n}(2n-1)!!(2n-7)(n-1)n(n+1)(7n^2-n+18)^2(n-3)^2(n-2)^2},$$

and

$$\langle v_4(n), v_4(n)\rangle = \tfrac{2048(n^2+13n-18)q(n)q_1(n)(n-1)!(2n-3)((2n-9)!!)^2}{2^{2n}(n+3)!(n-2)^3((n-3)(2n-3)!!(3n^3+28n^2-67n+28)(n^2+11n-6))^2}$$

with

$$q_1(n) = n^6 + 39n^5 + 445n^4 + 585n^3 - 7286n^2 + 9816n - 2880.$$

From (6.1) we see that we will need to borrow two functions in order to make an orthogonal intertwining MRA, and in order for these to be symmetric or antisymmetric we will set $w_{1,n} = \alpha_{1,0}(n)v_0 + \alpha_{1,2}(n)v_2 + v_4$ and $w_{2,n} = \alpha_{2,1}(n)u^1_{n,1} + u^1_{n,3}$. Taking note of the sign structure in (6.2), (5.5), and (6.6), (6.1) becomes the three equations

$$\begin{aligned}
|\langle r_{n,0}, l_{n,0}\rangle| &= \langle r_{n,0}, w_{1,n}\rangle^2 - \langle r_{n,0}, w_{2,n}\rangle^2,\\
0 &= \langle r_{n,0}, w_{1,n}\rangle\langle w_{1,n}, r_{n,1}\rangle - \langle r_{n,0}, w_{2,n}\rangle\langle w_{2,n}, r_{n,1}\rangle,\\
|\langle r_{n,1}, l_{1,n}\rangle| &= -\langle r_{n,1}, w_{1,n}\rangle^2 + \langle r_{n,1}, w_{2,n}\rangle^2.
\end{aligned}$$

These can be solved to give

$$\begin{aligned}
\frac{\langle r_{n,1}, w_{1,n}\rangle}{\sqrt{\langle w_{1,n}, w_{1,n}\rangle}} &= \sqrt{|\langle r_{n,1}, l_{1,n}\rangle| - \frac{\langle r_{n,1}, w_{2,n}\rangle^2}{\langle w_{2,n}, w_{2,n}\rangle}},\\
\frac{\langle r_{n,0}, w_{1,n}\rangle}{\sqrt{\langle w_{1,n}, w_{1,n}\rangle}} &= \frac{\langle r_{n,0}, w_{2,n}\rangle\langle r_{n,1}, w_{2,n}\rangle}{\sqrt{|\langle r_{n,1}, l_{1,n}\rangle|\langle w_{2,n}, w_{2,n}\rangle^2 - \langle r_{n,1}, w_{1,n}\rangle^2\langle w_{2,n}, w_{2,n}\rangle}},\\
0 &= |\langle r_{n,1}, l_{1,n}\rangle|\,|\langle r_{n,1}, l_{1,n}\rangle|\,\langle w_{2,n}, w_{2,n}\rangle + |\langle r_{n,1}, l_{1,n}\rangle|\,\langle r_{n,0}, w_{2,n}\rangle^2\\
&\quad - |\langle r_{n,1}, l_{1,n}\rangle|\,\langle r_{n,1}, w_{2,n}\rangle^2. \qquad (6.10)
\end{aligned}$$

With the definition of w_2 the last of the above equations can be rewritten as

$$\begin{aligned}
0 = &\begin{vmatrix} \langle u^1_{n,1}, u^1_{n,1}\rangle & -\langle u^1_{n,1}, r_{n,0}\rangle & \langle u^1_{n,1}, r_{n,1}\rangle\\ \langle u^1_{n,1}, r_{n,0}\rangle & \langle r_{n,0}, r_{n,0}\rangle & 0\\ \langle u^1_{n,1}, r_{n,0}\rangle & 0 & \langle r_{n,1}, r_{n,1}\rangle \end{vmatrix}\alpha^2_{2,1}\\
&+ 2\begin{vmatrix} \langle u^1_{n,1}, u^1_{n,3}\rangle & -\langle u^1_{n,3}, r_{n,0}\rangle & \langle u^1_{n,3}, r_{n,1}\rangle\\ \langle u^1_{n,1}, r_{n,0}\rangle & \langle r_{n,0}, r_{n,0}\rangle & 0\\ \langle u^1_{n,1}, r_{n,1}\rangle & 0 & \langle r_{n,1}, r_{n,1}\rangle \end{vmatrix}\alpha_{2,1}\\
&+ \begin{vmatrix} \langle u^1_{n,3}, u^1_{n,3}\rangle & -\langle u^1_{n,3}, r_{n,0}\rangle & \langle u^1_{n,3}, r_{n,1}\rangle\\ \langle u^1_{n,3}, r_{n,0}\rangle & \langle r_{n,0}, r_{n,0}\rangle & 0\\ \langle u^1_{n,3}, r_{n,1}\rangle & 0 & \langle r_{n,1}, r_{n,1}\rangle \end{vmatrix}.
\end{aligned}$$

Given the values of the inner products appearing in this equation, we find that either of the two values

$$\alpha_{2,1}(n) = \frac{2}{(n-1)(2n-3)}\left\{\frac{(2n-1)(3n-8)}{(n-2)} \pm 2\sqrt{7}\sqrt{\frac{(2n-1)(n+2)}{(2n-5)(n-2)}}\right\} \tag{6.11}$$

will suffice for $\alpha_{2,1}$. The first equation in (6.10) can be used to eliminate $\sqrt{\langle w_{1,n}, w_{1,n}\rangle}$ in the middle equation, and we have

$$\alpha_{1,2}(n) = \frac{(7n^2-n-18)\left\{q_3(n)(n-1)(n+1)(n+2)+q(n)\sqrt{7}\sqrt{(2n-1)(2n-5)(n+2)(n-2)}\right\}}{108(2n-5)(n+2)(n^2+11n-6)(3n^3+28n^2-67n+28)(n^4-6n^3-31n^2-22n+72)} \tag{6.12}$$

with

$$q_3(n) = 152n^6 + 3893n^5 + 19240n^4 - 76625n^3 - 105912n^2 + 330372n - 129600.$$

Now we solve for $\alpha_{1,0}(n)$ to find

$$\alpha_{1,0}(n) = \frac{\sqrt{10}}{108}\frac{q(n)\left\{\frac{(n-4)(n-2)(2n+1)}{n(n+2)(2n-7)}\left(q_4(n)+q_5(n)\sqrt{7}\sqrt{(2n-1)(2n-5)(n+2)(n-2)}\right)\right\}^{1/2}}{(2n-5)(n^2+11n-6)(3n^3+28n^2-67n+28)(n^4-6n^3-31n^2-22n+27)} \tag{6.13}$$

with

$$q_4(n) = 62n^6 + 2n^5 + 1480n^4 - 4526n^3 - 2250n^2 + 9774n - 5508$$

and

$$q_5(n) = 8n^4 + 95n^3 - 126n^2 + 297n - 162.$$

Knowing w_1 and w_2, we are now able to construct the orthogonal C^1 scaling functions. Let

$$h_0(t) = \begin{cases} 2|t|^3 - 3|t|^2 + 1 & \text{if } t \in [-1,1), \\ 0 & \text{otherwise} \end{cases}$$

and

$$h_1(t) = \begin{cases} (1-|t|)^2 t & \text{if } t \in [-1,1), \\ 0 & \text{otherwise.} \end{cases}$$

For $j = 4, \ldots, n$, set

$$\tilde{\phi}^j(\cdot) = \begin{cases} \phi_j^1(2\cdot - 1) & \text{if } t \in [0,1), \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{\phi}^2(\cdot) = \begin{cases} w_1(2\cdot - 1) & \text{if } t \in [0,1), \\ 0 & \text{otherwise,} \end{cases}$$

$$\tilde{\phi}^3(\cdot) = \begin{cases} w_2(2\cdot - 1) & \text{if } t \in [0,1), \\ 0 & \text{otherwise,} \end{cases}$$

and $\tilde{\phi}^i = (I - P_{\{\tilde{\phi}^2,\ldots,\tilde{\phi}^n,\tilde{\phi}^2(\cdot+1),\ldots,\tilde{\phi}^n(\cdot+1)\}})h_i$, $i = 0, 1$. Then the above computations give the following theorem.

THEOREM 6.1. *For $n \geq 6$, $\alpha_{1,0}(n)$, $\alpha_{1,2}(n)$, and $\alpha_{2,1}(n)$ given by (5.9), $\tilde{\Phi} = \{\tilde{\phi}^0, \ldots, \tilde{\phi}^n\}^*$ generates an orthogonal multiresolution analysis $\{\tilde{V}_k^{n,1}\}$. Furthermore, the last $n-1$ functions are symmetric or antisymmetric about $\frac{1}{2}$. The first function $\tilde{\phi}^0$ is symmetric about 0 while $\tilde{\phi}^1$ is antisymmetric about 0.*

Proof. Since h_0 and h_1 are linear combinations of similarly scaled versions of $r_j^1,\ j = 0, 1$ and $l_j^1,\ j = 0, 1$ the result follows from the computations above. □

With the scaling functions above we construct the coefficients $C_{n,i}^1,\ i = -2, -1, 0, 1$, in the matrix refinement equation. Theorem 4.4 implies that there will be four wavelets not supported in $[0, 1]$. Using arguments similar to Corollary 5.2 leads to the following.

COROLLARY 6.2. *Suppose $\tilde{\psi}^0, \ldots, \tilde{\psi}^3$ are chosen so that*

$$\begin{aligned}
\tilde{\psi}^0 &= \sqrt{2}(I - P_{\tilde{\phi}^0})\tilde{\phi}_{1,0}^0,\\
\tilde{\psi}^1 &\propto (I - P_{\{\tilde{\psi}^0, \tilde{\phi}^0\}})(\chi_{[0,1]} - \chi_{[-1,0]})(I - P_{\tilde{\phi}_{1,0}^1})\tilde{\phi}^1,\\
\tilde{\psi}^2 &\propto (I - P_{\{\tilde{\psi}^3, \tilde{\phi}^1\}})(\chi_{[0,1]} - \chi_{[-1,0]})(I - P_{\tilde{\phi}_{1,0}^0})\tilde{\phi}^0,\\
\tilde{\psi}^3 &= \frac{2\sqrt{2}}{\sqrt{7}}(I - P_{\tilde{\phi}^1})\tilde{\phi}_{1,0}^1
\end{aligned}$$

and $\tilde{\psi}^4, \ldots, \tilde{\psi}^n$ form a basis for Ψ^0 consisting only of functions symmetrical or antisymmetrical about $\frac{1}{2}$. Then $\{\tilde{\psi}^0, \ldots, \tilde{\psi}^n\}$ generates a shift-invariant orthonormal basis for W_0. Furthermore, $\tilde{\psi}^0(0) = \tilde{\phi}^0(0)$, $(\tilde{\psi}^3)'(0) = \sqrt{7}(\tilde{\phi}^1)'(0)$, $\tilde{\psi}^1(0) = 0$, and $(\tilde{\psi}^2)'(0) = 0$.

An example is given in Figure 2. In Figure 3 explicit formulas for these functions, accurate at least to 10 digits, are given. The coefficients in the refinement equation may be calculated using inner products. The wavelets supported on $[-1, 1]$ can be computed using Corollary 6.2. The wavelets supported in $[0, 1]$ can be obtained by finding $n-3$ orthogonal set of functions, symmetrical or antisymmetrical with respect to $\frac{1}{2}$, from the space $(I - P_{\{\tilde{\phi}^4, \ldots, \tilde{\phi}^{n-1}\}})A_1^{n,1} \ominus A_0^{n,1}$.

Theorem 4.5 shows that in order to construct a wavelet basis for $[0, 1]$ we need to rotate the scaling functions so that $\{\chi_{[0,1]}\tilde{\phi}^i\}_{i=0,1}$ is an orthogonal set as well as $\{\chi_{[-1,0]}\tilde{\phi}^i\}_{i=0,1}$. Exploiting the symmetry of $\tilde{\phi}^0$ and $\tilde{\phi}^1$, we find that $\hat{\phi}^0 = -\frac{1}{\sqrt{2}}\tilde{\phi}^0 - \frac{1}{\sqrt{2}}\tilde{\phi}^1$ and $\hat{\phi}^1 = -\frac{1}{\sqrt{2}}\tilde{\phi}^0 + \frac{1}{\sqrt{2}}\tilde{\phi}^1$ have the desired property. Let $\bar{\phi}_{k,j}^i = \hat{\phi}_{k,j}^i|_{[0,1]}$, $\bar{\psi}_{k,j}^m = \hat{\psi}_{k,j}^m|_{[0,1]}$, $m \neq 1, 3$, and for $m = 1, 3$ set

$$\bar{\psi}_{k,j}^m = \begin{cases} 0 & \text{if } \operatorname{supp} \hat{\psi}_{k,j}^m \cap [0,1]^c \neq \emptyset,\\ \hat{\psi}_{k,j}^m & \text{otherwise.} \end{cases}$$

Then from Theorem 4.5 we have the following theorem.

THEOREM 6.3. *The set $\{\bar{\phi}_{k,j}^i\colon\ k \geq 0,\ i = 1, \ldots, n,\ 0 \leq j \leq 2^k - 1 + \chi_{\{0,1\}}(i)\}$ is an orthogonal basis for $\bar{V}_k^{n,1} = \tilde{V}_k^{n,1} \cap L^2[0, 1]$ while $\{\bar{\psi}_{k,j}^i\colon\ k \geq 0,\ i = 1, \ldots, n,\ \chi_{\{1,3\}}(i) \leq j \leq 2^k - 1 + \chi_{\{0,2\}}(i)\}$ forms an orthogonal basis for $\bar{W}_k^{n,1} = \tilde{W}_k^{n,1} \cap L^2[0, 1]$. Furthermore, $\operatorname{cl}_{L^2} \bar{V}_0^{n,1} \oplus \bigoplus_{k \geq 0} \bar{W}_k^{n,1} = L^2[0, 1]$.*

Examples of these functions for $n = 6$ can be found in Figure 4.[2]

[2]Wavelets as well as the matrices in the refinement equation for this and other examples may be found at the web site www.math.gatech.edu/∼geronimo.

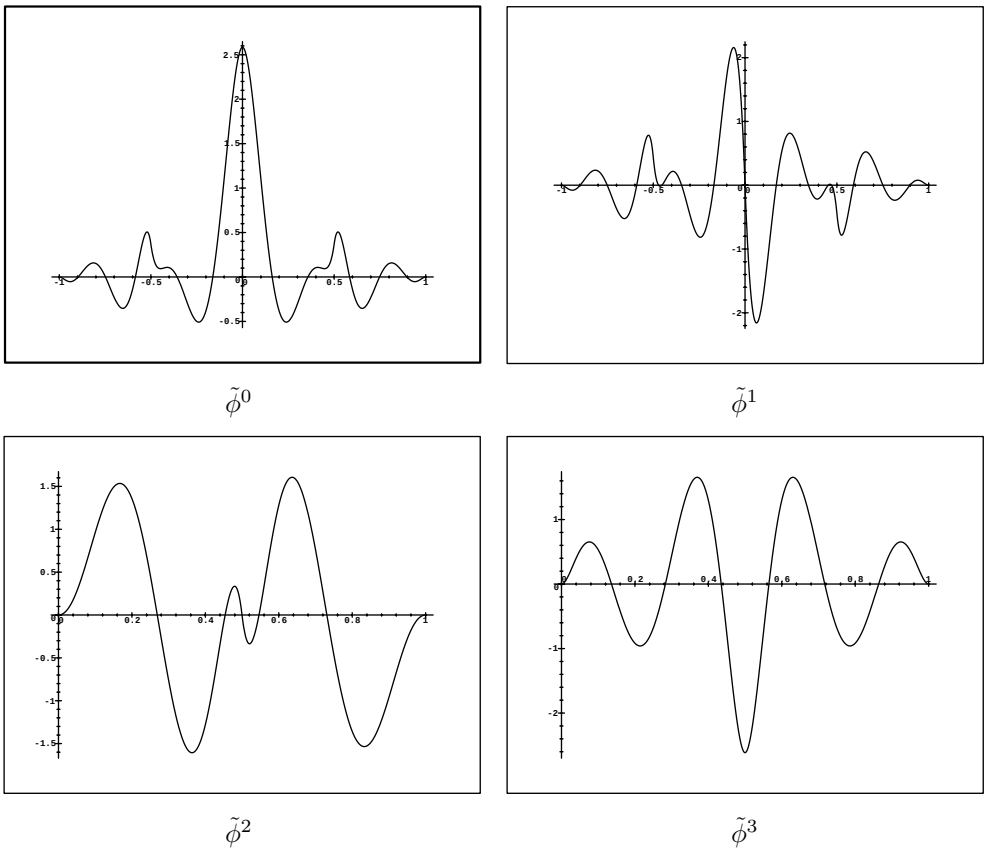


FIG. 2. *Selected C^1 scaling functions with approximation order* 7 ($n = 6$).

7. A C^2 example. The methods of the previous sections can be used to construct C^2 multiwavelets as well, although the formulas become extremely complicated. We will therefore content ourselves with briefly describing the procedure and exhibiting an example which may be of use. We do not prove that the procedure works for arbitrary n; however, we have verified it for a number of cases. For the C^2 case $k = 2$, $r_i^2 = (1+t)^3(1-t)^i$ and $l_i^2(t) = (1-t)^3(1+t)^i$, $i = 0, 1, 2$. The spaces $C_0^{n,2}(\frac{\cdot+1}{2})$ and $C_1^{n,2}(\frac{\cdot+1}{2})$ are each three dimensional. Hence we search for three orthonormal functions $w_1, w_2, w_3 \in A_1^{n,2}(\frac{\cdot+1}{2}) \ominus A_0^{n,2}(\frac{\cdot+1}{2})$ such that

$$\langle (I - P_{\{w_1,w_2,w_3\}}) r_i^{n,2}, (I - P_{\{w_1,w_2,w_3\}}) l_j^{n,2} \rangle = 0, \ i \le j = 0, 1, 2. \tag{7.1}$$

The above equation yields six nonlinear equations and in order to ease the computation somewhat we impose that $\langle w_3, r_0^{n,2} \rangle = 0$. By examining (7.1) we find that w_1 must satisfy four equations, w_3 five equations, and w_2 two equations. Thus, we choose $w_1 = a_{1,0}u_{n,0}^2 + a_{1,2}u_{n,2}^2 + a_{1,4}u_{n,4}^2 + a_{1,6}u_{n,6}^2$, $w_3 = a_{3,0}u_{n,0}^2 + a_{3,2}u_{n,2}^2 + a_{3,4}u_{n,4}^2 + a_{3,6}u_{n,6}^2 + a_{3,8}u_{n,8}^2$, and $w_2 = a_{2,1}u_{n,1}^2 + a_{2,3}u_{n,3}^2$ and observe that, as a consequence, n must be at least 11 for these functions all to have C^2 smoothness. For $n = 11$, we used Maple to obtain the following solution to 60 digits of accuracy:

$$\tilde{\phi}^0(t) = \begin{cases} 4067.904397t^6 - 3085.517213t^5 - 739.5537604t^4 \\ \quad +1113.129531t^3 - 249.0932934t^2 + 2.585173201 & \text{for } 0 \le t \le \frac{1}{2} \\ (-10946.18252t^4 + 29698.96665t^3 - 29673.11423t^2 \\ \quad +12935.83259t - 2076.394058)(t-1)^2 & \text{for } \frac{1}{2} < t \le 1 \\ \tilde{\phi}^0(-t) & \text{for } -1 \le t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^1(t) = \begin{cases} t(-12435.14749t^5 + 11555.74525t^4 - 175.1583524t^3 \\ \quad -2704.493962t^2 + 885.9596767t - 79.47577362) & \text{for } 0 \le t \le \frac{1}{2} \\ (17496.63084t^4 - 47730.65634t^3 + 47977.37771t^2 \\ \quad -21054.27094t + 3403.826649)(t-1)^2 & \text{for } \frac{1}{2} < t \le 1 \\ -\tilde{\phi}^1(-t) & \text{for } -1 \le t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^2(t) = \begin{cases} (2t-1)t^2(-12193.17741t^3 + 8033.232335t^2 \\ \quad -923.0732055t - 95.73716085) & \text{for } 0 \le t \le \frac{1}{2} \\ -\tilde{\phi}^2(1-t) & \text{for } \frac{1}{2} < t \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

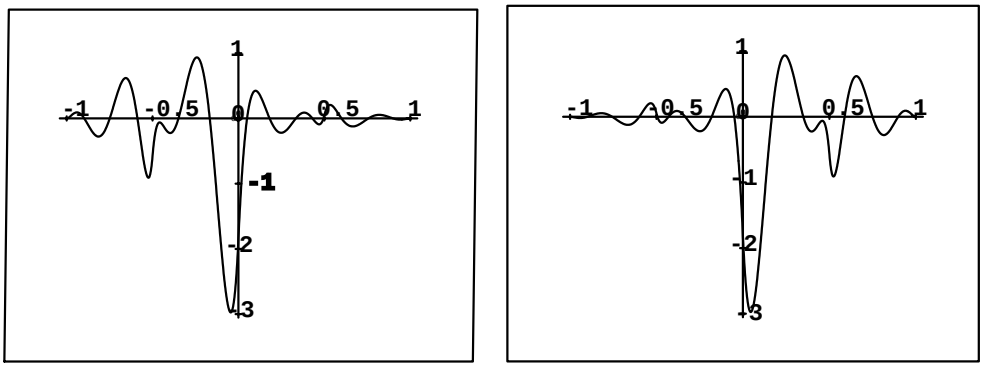
$$\tilde{\phi}^3(t) = \begin{cases} (54655.48659t^4 - 76071.45058t^3 + 37177.81845t^2 \\ \quad -7410.133638t + 493.1155758)t^2 & \text{for } 0 \le t \le \frac{1}{2} \\ \tilde{\phi}^3(1-t) & \text{for } \frac{1}{2} < t \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^4(t) = 3\sqrt{70}t^2(t-1)^2\chi_{[0,1]}$$

$$\tilde{\phi}^5(t) = 3\sqrt{770}t^2(2t-1)(t-1)^2\chi_{[0,1]}$$

$$\tilde{\phi}^6(t) = 3\sqrt{182}t^2(22t^2 - 22t + 5)(t-1)^2\chi_{[0,1]}$$

FIG. 3. *Formulas for the C^1 scaling functions.*



$\hat{\phi}^0$ $\hat{\phi}^1$

FIG. 4. *Modified C^1 scaling functions of degree 6 for truncation to $[0,1]$.*

$$\tilde{\phi}^0(t) = \begin{cases} 960.8390712t^{11} - 5925.739238t^{10} + 13283.80990t^9 \\ \quad - 11118.85044t^8 - 4083.423604t^7 + 15463.88643t^6 \\ -12421.39021t^5 + 4470.423870t^4 - 629.0433748 * t^3 - 1.5t^2 + 1 & \text{for } 0 \le t < \frac{1}{2} \\ (1-t)^3(-1188.938305t^8 + 3170.302415t^7 - 3222.784994t^6 \\ \quad + 1570.468582t^5 - 377.8071111t^4 + 37.83633397t^3 \\ \quad + 1.585938501t^2 - .6720319852t + .01241619153) & \text{for } \frac{1}{2} < t \le 1 \\ \tilde{\phi}^0(-t) & \text{for } -1 \le t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^1(t) = \begin{cases} t(-43.55084884t^{10} + 268.8737383t^9 - 588.5150956t^8 \\ \quad + 426.0982641t^7 + 364.8815486t^6 - 923.7069266t^5 \\ + 722.0921528t^4 - 265.9177149t^3 + 39.49465380t^2 + .75t - .5) & \text{for } 0 \le t < \frac{1}{2} \\ (1-t)^3(55.58117958t^8 - 152.6750597t^7 + 161.0387487t^6 \\ \quad - 81.78617419t^5 + 20.22050034t^4 - 1.881383443t^3 \\ \quad - .1197243402t^2 + .03075948978t - .0002283505048) & \text{for } \frac{1}{2} \le t \le 1 \\ -\tilde{\phi}^1(-t) & \text{for } -1 \le t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^2(t) = \begin{cases} -4.290797794t^{11} + 25.45780265t^{10} - 49.23874622t^9 \\ + 10.99949461t^8 + 96.73006250t^7 - 163.9201834t^6 + 126.4483297t^5 \\ - 52.49395898t^4 + 11.30900992t^3 - 1.002944440t^2 + .001962959931 & \text{for } 0 \le t < \frac{1}{2} \\ (1-t)^3(5.245026448t^8 - 14.95000551t^7 + 16.41738822t^6 \\ \quad - 8.639953559t^5 + 2.137676365t^4 - .1611170142t^3 \\ \quad - .02104722553t^2 + .002877121114t + .00003155317905) & \text{for } \frac{1}{2} < t \le 1 \\ \tilde{\phi}^0(-t) & \text{for } -1 \le t < 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^3(t) = \begin{cases} t^3(837.0653712t^8 - 1385.345634t^7 + 2498.141505t^6 \\ - 19046.31150t^5 + 51025.24775t^4 - 61978.36348t^3 \\ \quad + 38423.77644t^2 - 11768.03455t + 1393.219445) & \text{for } 0 \le t \le \frac{1}{2} \\ \tilde{\phi}^3(1-t) & \text{for } \frac{1}{2} < t \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^4(t) = \begin{cases} (13553.28638t^7 - 72947.95448t^6 + 155109.9841t^5 \\ - 159752.4874t^4 + 73088.74021t^3 - 2106.644683t^2 \\ \quad - 8799.858303t + 1848.967175)t^3(t - 1/2) & \text{for } 0 \le t \le \frac{1}{2} \\ -\tilde{\phi}^4(1-t) & \text{for } \frac{1}{2} < t \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^5(t) = \begin{cases} t^3(19651.08243t^8 - 58017.27598t^7 + 17880.79147t^6 \\ + 126264.0351t^5 - 207975.2506t^4 + 145513.9138t^3 \\ \quad - 51576.42468t^2 + 8813.779535t - 552.7580636) & \text{for } 0 \le t \le \frac{1}{2} \\ \tilde{\phi}^5(1-t) & \text{for } \frac{1}{2} < t \le 1 \\ 0 & \text{elsewhere} \end{cases}$$

$$\tilde{\phi}^6(t) = \sqrt{6006}t^3(1-t)^3\chi_{[0,1]}$$

$$\tilde{\phi}^7(t) = \sqrt{30030}t^2(2t-1)(1-t)^3\chi_{[0,1]}$$

$$\tilde{\phi}^8(t) = 2\sqrt{7293}t^3(30t^2 - 30t + 7)(1-t)^3\chi_{[0,1]}$$

$$\tilde{\phi}^9(t) = 2\sqrt{40755}t^3(2t-1)(34t^2 - 34t + 7)(1-t)^3\chi_{[0,1]}$$

FIG. 5. *Formulas for the scaling functions.*

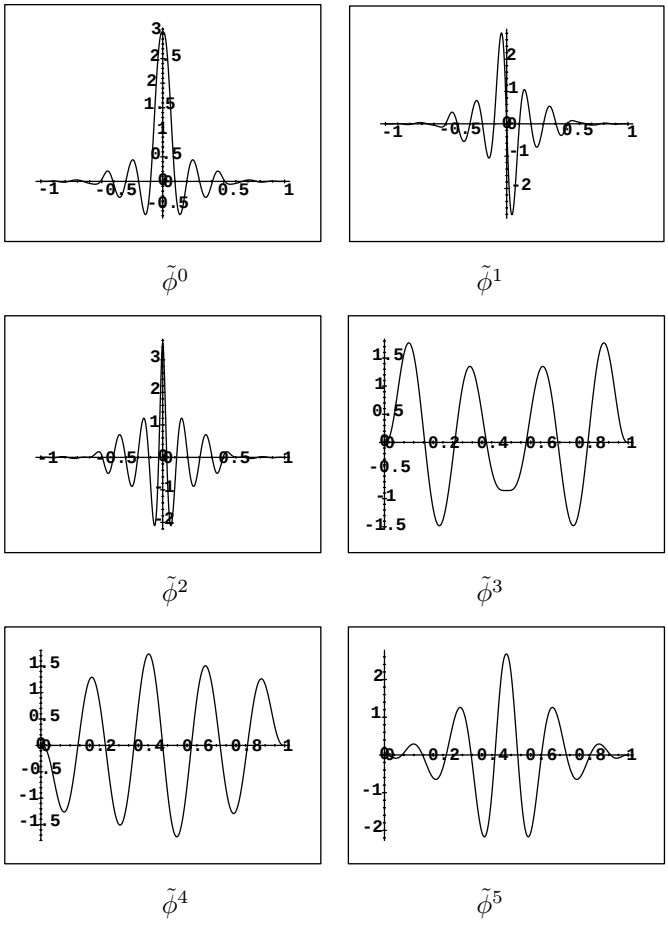


FIG. 6. *Selected C^2 scaling functions with approximation order* 12 ($n = 11$).

$$a_{1,0} = 837.06537121043762613176899215628138472454578263465164388 3901,$$
$$a_{1,2} = 34683.28058531315475474840518015915154434934560041504686163 72,$$
$$a_{1,4} = 98577.94641657184834001461016465742561305738952570617373444 47,$$
$$a_{1,6} = 23873.28589415915260733303499008456445444578751981672251759 43,$$
$$a_{2,1} = 69361.55251081876097926389614291036929402325065142922835779 45,$$
$$a_{2,3} = 135633.53024617074706623955211072817787004123086530793824906 7,$$
$$a_{3,0} = 19651.08242733520672722369351521421437937555480828400437002 10,$$
$$a_{3,2} = 518517.56518403363191974586932873252940671472987789656146002 5,$$
$$a_{3,4} = 968629.60666757961957220956120267186030658034982386738309165 7,$$
$$a_{3,6} = 236239.02079854637651178042421251670380169476066182744618324 0,$$
$$a_{3,8} = 3210.54994458080183556006870726795783137235163979326960039700.$$

The corresponding functions are given with 10 digits of accuracy in Figure 5, and their graphs are shown in Figure 6. The orthonormal wavelets can be calculated using techniques similar to those of sections 5 and 6. One final Gram–Schmidt step is necessary since two of the three functions in $(I - P_{Q_0})Q_1$ are symmetric. The same is true of $(I - P_{Q_0 \cup \Psi^s})Z$.

Acknowledgment. J. S. Geronimo would like to thank the members of the Theoretical Physics Division at Saclay and the Laboratoire d'Analyse Numerique at the University of Pierre and Marie Curie for their hospitality and support during the time this work was being completed.

REFERENCES

[1] B. Alpert, *Sparse Representation of Smooth Linear Operators*, Ph.D. thesis, Yale University, New Haven, CT, 1990.
[2] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, PA, 1992.
[3] W.N. Bailey, *Generalized Hypergeometric Series*, Cambridge Tracts in Math. Math. Phys. 32, Cambridge University Press, Cambridge, UK, 1964.
[4] G.C. Donovan, *Fractal Functions, Splines, and Wavelets*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA, 1995.
[5] G.C. Donovan, J.S. Geronimo, and D.P. Hardin, *Intertwining multiresolution analyses and the construction of piecewise polynomial wavelets*, SIAM J. Math. Anal., 27 (1996), pp. 1791–1815.
[6] G.C. Donovan, J.S. Geronimo, and D.P. Hardin, *Fractal functions, splines, intertwining multiresolution analyses and wavelets*, in Proceedings SPIE, vol. 2303, A.F. Laine and M.A. Unser, eds., 1994, pp. 238–243.
[7] G.C. Donovan, J.S. Geronimo, and D.P. Hardin, *C^0 spline wavelets with arbitrary approximation order*, in Proceedings SPIE, vol. 2569, A.F. Laine, M.A. Unser, and M.V. Wickerhauser, eds., 1995, pp. 376–380.
[8] G.C. Donovan, J.S. Geronimo, and D.P. Hardin, *Families of compactly supported orthogonal spline wavelets*, in Proceedings, International Conference on Scientific Computing and Modeling, 1995, to appear.
[9] G.C. Donovan, J.S. Geronimo, D.P. Hardin, and P.R. Massopust, *Construction of orthogonal wavelets using fractal interpolation functions*, SIAM J. Math. Anal., 27 (1996), pp. 1158–1192.
[10] B. Fischer and J. Prestin, *Wavelets based on orthogonal polynomials*, Math. Comp., 66 (1997), pp. 1593–1618.
[11] T.N.T. Goodman and S.L. Lee, *Wavelets of multiplicity r*, Trans. Amer. Math. Soc., 342 (1994), pp. 307–324.
[12] I.S. Gradshteyn and I.M. Ryzhik, *Tables of Integrals, Series, and Products*, Academic Press, New York, 1965.
[13] D.P. Hardin, B. Kessler, and P.R. Massopust, *Multiresolution analyses and fractal functions*, J. Approx. Theory, 71 (1992), pp. 104–120.
[14] L. Herve, *Analyses multirésolutions de multiplicité d. applications à l'interpolation dyadique*, Appl. Comput. Harmonic Anal., 1 (1994), pp. 299–315.
[15] T. Kilgore and J. Prestin, *Polynomial wavelets on an interval*, Constr. Approx., 12 (1996), pp. 95–110.
[16] S. Mallat, *Multiresolution approximations and wavelet orthonormal bases of $L^2(\mathbf{R})$*, Trans. Amer. Math. Soc., 315 (1994), pp. 69–87.
[17] Y. Meyer, *Ondelettes et Opérateurs*, Hermann, Paris, 1990.
[18] G. Plonka, K. Selig, and M. Tasche, *On the construction of wavelets on a bounded interval*, Adv. Comput. Math., 3 (1995), pp. 1–14.
[19] G. Strang and V. Strela, *Finite element wavelets*, in Proceedings SPIE, vol. 2303, A.F. Laine and M.A. Unser, eds., 1994, pp. 202–213.
[20] G. Szegö, *Orthogonal Polynomials*, Amer. Math. Soc. Colloq. Pub. 23, AMS, Providence, RI, 1939.
[21] J.A. Wilson, *Three-term contiguous relations and some new orthogonal polynomials*, in Pade and Rational Approximation, Theory and Applications, E.B. Saff and R.S. Varga, eds., Academic Press, New York, 1977.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1057–1075

INFINITE PRODUCTS AND NORMALIZED QUOTIENTS OF HYPERGEOMETRIC FUNCTIONS*

S.-L. QIU† AND M. VUORINEN‡

Abstract. For $r \in (0,1)$ and $a \in (0,1)$ the authors consider the quotient of hypergeometric functions

$$\mu_a(r) \equiv cF(a,1-a;1;1-r^2)/F(a,1-a;1;r^2),$$

where the normalizing coefficient $c = \pi/(2\sin(\pi a))$. With this choice of c, $\mu(r) \equiv \mu_{1/2}(r)$, where $\mu(r)$ is the modulus of the Grötzsch ring $B^2 \setminus [0,r]$ in the plane. A new infinite product expansion is given for $\mu(r)$. It is shown that several well-known properties of the function $\mu(r)$ have their counterparts for $\mu_a(r)$.

Key words. zero-balanced hypergeometric functions, modulus of Grötzsch ring, infinite product

AMS subject classifications. Primary, 30C62, 33C05; Secondary, 26D15, 11F99

PII. S0036141097326805

1. Introduction.

As usual, for real numbers a, b, and c with $c \neq 0, -1, -2, \ldots$, let

$$F(a,b;c;x) := {}_2F_1(a,b;c;x) \equiv \sum_{n=0}^{\infty} \frac{(a,n)(b,n)}{(c,n)} \frac{x^n}{n!} \tag{1.1}$$

for $x \in (-1,1)$ denote the *Gaussian hypergeometric function* [AS], [Ask1], [R]. Here $(a,0) = 1$ for $a \neq 0$ and (a,n) is the shifted factorial function

$$(a,n) := a(a+1)(a+2)\cdots(a+n-1)$$

for $n \in \mathbf{N} \equiv \{k : \ k \text{ is a positive integer}\}$. The function $F(a,b;c;x)$ is said to be *zero balanced* if $c = a+b$. It is well known that $F(a,b;c;x)$ has many important applications, and many classes of special functions in mathematical physics are particular or limiting cases of this function. For these and for properties of $F(a,b;c;x)$, see, for example, [AS], [Ao], [Ask1], [Ask2], [Be1], [Be2], [Be3], [Be4], [CC], [E], [R], [Var], [Va], [WW], and [WZ]. Here we recall that only in the special cases when $a = b = 1/2$ and $-a = b = 1/2$, we have for $x \in (0,1)$ and $x' = \sqrt{1-x^2}$,

$$\mathcal{K}(x) \equiv \frac{\pi}{2} F\Big(\frac{1}{2}, \frac{1}{2}; 1; x^2\Big) = \int_0^{\pi/2} (1 - x^2 \sin^2 t)^{-1/2} dt, \quad \mathcal{K}'(x) \equiv \mathcal{K}(x'), \tag{1.2}$$

and

$$\mathcal{E}(x) = \frac{\pi}{2} F\Big(-\frac{1}{2}, \frac{1}{2}; 1; x^2\Big) = \int_0^{\pi/2} (1 - x^2 \sin^2 t)^{1/2} dt, \quad \mathcal{E}'(x) \equiv \mathcal{E}(x'), \tag{1.3}$$

*Received by the editors September 3, 1997; accepted for publication (in revised form) September 8, 1998; published electronically August 16, 1999.
http://www.siam.org/journals/sima/30-5/32680.html

†President's Office, Hangzhou Institute of Electronics Engineering, Hangzhou 310037, People's Republic of China.
‡Department of Mathematics, P.O. Box 4 (Yliopistonkatu 5), University of Helsinki, FIN-00100 Helsinki, Finland (vuorinen@csc.fi).

which are known as the *complete elliptic integrals of the first kind and of the second kind*, respectively [Bo], [BF].

Let B^2 be the unit disk in the plane and $\mu(r)$ the *modulus of the plane Grötzsch ring* $B^2 \setminus [0,r]$ for $r \in (0,1)$ [LV]. It is well known that $\mu(r)$ has the following explicit expression [LV, p. 60]:

$$\mu(r) = \frac{\pi}{2}\frac{\mathcal{K}'(r)}{\mathcal{K}(r)}, \quad \text{for} \quad 0 < r < 1 \quad \text{and} \quad r' = \sqrt{1-r^2}. \tag{1.4}$$

The special function $\mu(r)$ plays a very important role in geometric function theory, quasi-conformal theory, and quasi-regular theory (cf. [AVV2], [LV], and [Vu]). It also has applications in some other mathematical fields such as the theory of analytic functions and number theory. In number theory, for example, it appears in the *classical modular equation of signature* 2 *and degree* p, $p > 1$, i.e., the equation

$$\mu(s) = p\mu(r), \quad 0 < r < 1,$$

(see [Be3] and [BB]), while in the theory of analytic functions, by [M, Theorem 1.1], the Schottky upper bound can be expressed in terms of $\mu(r)$. Numerous properties of $\mu(r)$ have been obtained (see, for instance, [AV], [AVV2], and [LV]).

A natural generalization of $\mu(r)$ is the homeomorphism $\mu_a : (0,1) \to (0,\infty)$ defined by

$$\mu_a(r) \equiv \frac{\pi}{2\sin\pi a}\frac{F(a,1-a;1;1-r^2)}{F(a,1-a;1;r^2)} \tag{1.5}$$

for $a, r \in (0,1)$. Clearly, $\mu_{1-a}(r) = \mu_a(r)$. Hence, we may assume that $0 < a \le 1/2$. Using the function $\mu_a(r)$, one can write the so-called *generalized modular equation of signature* $1/a$ *and degree* p, $p > 1$, as

$$\mu_a(s) = p\mu_a(r), \quad 0 < r < 1, \tag{1.6}$$

which was studied by S. Ramanujan (cf. [Ask2] and [Be2]), and was recently studied in [BBG], [Be5], and [Ga]. In particular, several beautiful identities satisfied by r and s were obtained in [BBG]. Properties of $\mu_a(r)$, of course, are indispensable in the study of (1.6).

Since $\mu_{1/2}(r) \equiv \mu(r)$, and since as a function of the parameter a, $\mu_a(r)$ is analytic on $(0,1/2]$, it is natural to ask whether every known result for $\mu(r)$ has a counterpart with $\mu(r)$ replaced by $\mu_a(r)$, and how to extend the well-known results for $\mu(r)$ to the function $\mu_a(r)$.

On the other hand, it is well known that [J, p. 146]

$$\exp(\mu(r) + \log r) = 4\prod_{n=1}^{\infty}\left(\frac{1+q^{2n}}{1+q^{2n-1}}\right)^4 \tag{1.7}$$

for $r \in (0,1)$, where $q = \exp(-2\mu(r))$. By virtue of (1.7), several infinite-product representations and inequalities have been obtained for $\mu^{-1}(x)$ and for the Hersch–Pfluger φ-distortion function [HP]

$$\begin{cases} \varphi_K(r) \equiv \mu^{-1}(\mu(r)/K), \\ \varphi_K(0) = \varphi_K(1) - 1 = 0 \end{cases} \tag{1.8}$$

for $r \in (0,1)$ and $K \in (0,\infty)$ (see [AV], [AVV2], and [VV]). However, since the Jacobi product on the right side of (1.7) involves $\mu(r)$, it is not very convenient for one to employ (1.7) to do numerical computation for $\mu(r)$ and to study some related special functions such as $\varphi_K(r)$.

One purpose of the present paper is to find a new infinite-product representation for $\mu(r)$ which involves only r, and to extend this representation to the function $\mu_a(r)$. We shall prove the following result.

THEOREM 1.1. *For $a \in (0,1/2]$, let $R(a) \equiv R(a,1-a)$, where $R(a,b)$ is defined by (2.5), and for $r \in (0,1), n \in \mathbf{N}$, let $r_0 = r' = \sqrt{1-r^2}$, and*

$$r_1 = \varphi_2(r') = \frac{2\sqrt{r'}}{1+r'}, \ldots, r_n = \varphi_2(r_{n-1}) = \frac{2\sqrt{r_{n-1}}}{1+r_{n-1}} = \varphi_{2^n}(r'). \tag{1.9}$$

Then

$$\prod_{n=0}^{\infty}(1+r_n)^{2^{-n}} \le \exp(\mu_a(r)+\log r) \le \frac{1}{4}\exp(R(a)/2)\ \prod_{n=0}^{\infty}(1+r_n)^{2^{-n}}, \tag{1.10}$$

with equality for all $r \in (0,1)$ in each inequality if and only if $a = 1/2$. In particular, for all $r \in (0,1)$,

$$\exp(\mu(r)+\log r) = \prod_{n=0}^{\infty}(1+r_n)^{2^{-n}} \tag{1.11}$$

or equivalently,

$$\mu(r) = \log\frac{1}{r} + \sum_{n=0}^{\infty}\frac{1}{2^n}\log(1+r_n).$$

Another purpose of this paper is to extend some well-known properties of $\mu(r)$ to $\mu_a(r)$ and to show that some other properties of $\mu(r)$ cannot be extended to $\mu_a(r)$ for $a \in (0,1/2)$.

Throughout this paper, we let $r' = \sqrt{1-r^2}$ for $r \in [0,1]$, and let *arth* denote the inverse of the hyperbolic tangent *th* and $R(a)$ be as in Theorem 1.1.

We now state some of our other main results, which serve the second purpose of this paper.

THEOREM 1.2. *Let $a \in (0,1/2]$. Then we have the following:*

1. *The function $f(r) \equiv r'\mu_a(r)/\log(1/r)$ is strictly increasing from $(0,1)$ onto $(1,\infty)$.*
2. *The function $g(r) \equiv \mu_a(1/r)/\log r$ is strictly decreasing and convex from $(1,\infty)$ onto $(1,\infty)$. However, the function $G(r) \equiv g(1/r)$ is neither concave nor convex on $(0,1)$.*
3. *The function $h(r) \equiv \mu_a(r)/\{[R(a)/2]+\log(1/r)\}$ is strictly decreasing and concave from $(0,1)$ onto $(0,1)$.*
4. *The function $H(r) \equiv h(r)/\sqrt{r'}$ is strictly increasing from $(0,1)$ onto $(1,\infty)$. Moreover, for all $a \in (0,1/2]$ and $r \in (0,1)$,*

$$\left[\frac{1}{2}R(a)-\log r\right]\sqrt{r'} < \mu_a(r) < \frac{1}{2}R(a)-\log r. \tag{1.12}$$

THEOREM 1.3. *Let $a \in (0,1/2]$. Then we have the following:*

1. *The function* $f(r) \equiv \mu_a(e^{-r})$ *is strictly increasing and concave from* $(0,\infty)$ *onto* $(0,\infty)$. *In particular, for all* $a \in (0,1/2]$, *and* $x, y, p \in (0,1)$,

$$p\mu_a(x) + (1-p)\mu_a(y) \le \mu_a(x^p y^{1-p}) \tag{1.13}$$

and

$$\mu_a(x) + \mu_a(y) \le 2\mu_a(\sqrt{xy}). \tag{1.14}$$

Equality holds in (1.13) *and in* (1.14) *if and only if* $x = y$.

2. *The function* $g(r) \equiv \mu_a(1/r)$ *is strictly increasing and concave from* $(1,\infty)$ *onto* $(0,\infty)$. *In particular, for all* $a \in (0,1/2]$ *and* $x, y, p \in (0,1)$,

$$p\mu_a(x) + (1-p)\mu_a(y) \le \mu_a\left(\frac{xy}{(1-p)x + py}\right), \tag{1.15}$$

with equality if and only if $x = y$.

3. *For each* $t \in (0,1)$, *the function* $h(r) \equiv \mu_a(\frac{rt}{1+r't'}) - \mu_a(r)$ *is strictly increasing from* $(0,1)$ *onto* $(arth\, t', \mu_a(t))$. *Moreover, for all* $a \in (0,1/2]$ *and* $r, t \in (0,1)$,

$$\begin{cases} \mu_a(r) + \mu_a(t) - \dfrac{1}{2}[R(a) - \log 4] < \mu_a(r) + artht' \\ \qquad\qquad < \mu_a\left(\dfrac{rt}{1+r't'}\right) < \mu_a(r) + \mu_a(t). \end{cases} \tag{1.16}$$

4. *The function* $G(r) \equiv \mu_a(r)/\mu_a(\sqrt{r})$ *is strictly decreasing from* $(0,1)$ *onto* $(1,2)$. *In particular, for all* $a \in (0,1/2]$ *and* $r \in (0,1)$,

$$\mu_a(r) < \mu_a(r^2) < 2\mu_a(r). \tag{1.17}$$

THEOREM 1.4. *Let* $a \in (0,1/2]$. *Then the following apply:*

1. *Both of the functions* $f(r) \equiv \mu_a(r)$ *and* $1/f(r)$ *have exactly one inflection point on* $(0,1)$.
2. *The function* $g(r) \equiv \mu_a(r) + \log(r/r')$ *is strictly increasing and convex from* $(0,1)$ *onto* $(R(a)/2, \infty)$.
3. *The function* $h(r) \equiv r\exp(\mu_a(r))$ *is strictly decreasing and concave from* $(0,1)$ *onto* $(1, \exp(R(a)/2))$.
4. *Define the function* G *on* $(0,1)$ *by*

$$G(r) = \mu_a(r) + \log(r/\sqrt{r'}).$$

Then we have the following:

i. G *is strictly increasing on* $(0,1)$ *if and only if* $a = 1/2$.

ii. G *is convex on* $(0,1)$ *if and only if* $a = 1/2$.

iii. *If* $a \in (0,1/2)$, *then there exists a unique* $r_0 \in (0,1)$ *such that* G *is strictly decreasing on* $(0,r_0]$, *and increasing on* $[r_0,1)$, *with* $G(0^+) = R(a)/2$ *and* $G(1^-) = \infty$. *Moreover,* G *is neither concave nor convex on* $(0,1)$ *in this case.*

Remark 1. 1. It should be indicated that in addition to (1.7), there is a classical infinite-product representation [BB, Formula (2.5.15), p. 52]

$$\exp(\mu(r) + \log r) = 4\sqrt{r'} \prod_{n=1}^{\infty} \left(\frac{a_n}{b_n}\right)^{3/2^{n+1}} \tag{1.18}$$

for $r \in (0,1)$, where

$$a_0 = 1, \quad b_0 = r', \quad a_n = (a_{n-1} + b_{n-1})/2 \quad \text{and} \quad b_n = \sqrt{a_{n-1}b_{n-1}}$$

for $n \in \mathbf{N}$. However, as we can see, it is more convenient for one to use (1.11) to obtain some properties of $\mu(r)$ and some other related special functions such as $\varphi_K(r)$ than to use (1.18). For example, from (1.11) we get

$$\mu(r) - arth\ r' = \sum_{n=1}^{\infty} \frac{1}{2^n} \log(1 + r_n). \tag{1.19}$$

It is well known that as a function of r, $\varphi_K(r)$ is strictly concave on $(0,1)$ if $K > 1$ (cf. [AVV2, Exercises 10.18(9)]). Clearly, r' is concave on $(0,1)$, and $\log x$ is strictly increasing and concave on $(0,\infty)$. Hence, the right side of (1.19) is a sum of strictly decreasing and concave functions of r on $(0,1)$ (cf. [AQV, Lemma 2.1(1)]), so that as we know, $\mu(r) - arth\ r'$ is strictly decreasing and concave from $(0,1)$ onto $(0, \log 2)$.

2. In the special case when $a = 1/2$, the results in Theorems 1.2, 1.3, and 1.4 can be found in [AVV2] and in references therein.

3. Several known results for $\mu(r)$ have been generalized to $\mu_a(r)$ in [BPV] and [QVu2].

4. The inequality (1.14) has been proved by two different methods in [BPV, Theorem 1.5] and [QVu2, Theorem 1.18(1)]. In Theorem 1.3 item 2, we use a new method to obtain a more general inequality (1.13). We observe that (1.14) is the second part of [AVV1, Open Problem 10, p. 80].

5. The results in Theorems 1.1, 1.2, 1.3 and 1.4 enable us to show some properties of the solution of the generalized modular equation (1.6)

$$s = \varphi^a_{1/p}(r) \equiv \mu_a^{-1}(p\mu_a(r)). \tag{1.20}$$

These and some other applications will appear in a separate paper [AQVV].

2. Some properties of $F(a, b; c; x)$. In this section, we study some monotonicity properties of the function $F(a, b; c; x)$ and certain of its combinations. These are needed in the proofs of the main theorems stated in section 1. But first, we recall some known results for the function $F(a, b; c; x)$, which will be frequently used in what follows.

It is well known that the properties of the hypergeometric functions are closely related to those of the *gamma function* $\Gamma(x)$, the *psi function* $\Psi(x)$, and the *beta function* $B(x, y)$. For positive numbers x and y, these functions are defined by

$$\Gamma(x) = \int_0^{\infty} e^{-t} t^{x-1} dt, \quad \Psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}, \quad B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}, \tag{2.1}$$

respectively (cf. [WW]). It is well known that the gamma function satisfies the *difference equation* [WW, p. 237]

$$\Gamma(x+1) = x\Gamma(x) \tag{2.2}$$

if x is not a nonpositive integer and has the so-called *reflection property* [WW, p. 239]

$$\Gamma(x)\Gamma(1-x) = \frac{\pi}{\sin \pi x} = B(x, 1-x) \tag{2.3}$$

if x is not an integer. From (2.2), it follows that [AS, 6.3.6]

$$\Psi(x+n)=\sum_{k=1}^{n}\frac{1}{x+n-k}+\Psi(x) \tag{2.4}$$

for $n\in\mathbf{N}$. We shall also need the function

$$R(a,b)=-2\gamma-\Psi(a)-\Psi(b), R(1/2,1/2)=\log 16, \tag{2.5}$$

where γ is the *Euler–Mascheroni constant* defined by

$$\gamma=\lim_{n\to\infty}\left(\sum_{k=1}^{n}\frac{1}{k}-\log n\right)=0.577215\ldots. \tag{2.6}$$

By [QVu2, Lemma 2.14(2)], for $a\in(0,1/2]$,

$$R(a)\equiv R(a,1-a)\geq A[(1/2)-a]^2+\log 16, \tag{2.7}$$

with equality if and only if $a=1/2$, where $A=14\zeta(3)=16.82879\ldots$ and $\zeta(x)$ is the *Riemann zeta function*.

Two of the important tools we shall need in our work are Ramanujan's asymptotic formula (see [Ask1], [Be2], and [E])

$$B(a,b)F(a,b;a+b;r)+\log(1-r)=R(a,b)+\mathrm{O}((1-r)\log(1-r)) \tag{2.8}$$

(for $a,b\in(0,\infty)$) as r tends to 1, which is a special case of [AS, 15.3.10], and the following Ramanujan's derivative formula [Be2, Corollary, p. 86]

$$\frac{d}{dx}\left[\frac{F(a,1-a;1;1-x)}{F(a,1-a;1;x)}\right]=-\frac{\sin\pi a}{\pi x(1-x)F(a,1-a;1;x)^2} \tag{2.9}$$

for $a,x\in(0,1)$. From (2.9) we immediately get the derivative of $\mu_a(r)$ with respect to r: For $a,r\in(0,1)$,

$$\frac{d\mu_a(r)}{dr}=-\frac{1}{rr'^2F(a,1-a;1;r^2)^2}. \tag{2.10}$$

By (1.1) we have

$$\frac{d}{dx}F(a,b;c;x)=\frac{ab}{c}F(a+1,b+1;c+1;x). \tag{2.11}$$

It follows from (2.9) and (2.11) that

$$F(1+a,2-a;2;1-x)F(a,1-a;1;x) \tag{2.12}$$

$$+F(1+a,2-a;2;x)F(a,1-a;1;1-x)=\frac{\sin\pi a}{\pi a(1-a)x(1-x)}$$

for $a,x\in(0,1)$.

Other important tools in the rest of this paper are the following Landen inequalities.

THEOREM 2.1 (see [QVu1, Theorem 1.9(1)]). *For $a, b \in (0,1)$ with $c = a + b \le 1$ (< 1, respectively), the function*

$$f(r) \equiv (1+r)F(a,b;c;r^2) - F(a,b;c;4r/(1+r)^2)$$

is increasing (strictly, respectively) from $(0,1)$ onto $(0, [R(a,b) - \log 16]/B(a,b))$. In particular, for all $a, b, r \in (0,1)$ with $c = a + b \le 1$,

$$\begin{cases} F\Big(a,b;c;\Big(\dfrac{2\sqrt{r}}{1+r}\Big)^2\Big) \le (1+r)F(a,b;c;r^2) \\ \le F\Big(a,b;c;\Big(\dfrac{2\sqrt{r}}{1+r}\Big)^2\Big) + \dfrac{1}{B(a,b)}[R(a,b) - \log 16] \end{cases} \tag{2.13}$$

and

$$\begin{cases} \dfrac{1+r}{2}F(a,b;c;1-r^2) \le F\Big(a,b;c;\Big(\dfrac{1-r}{1+r}\Big)^2\Big) \\ \le \dfrac{1+r}{2}\Big\{F(a,b;c;1-r^2) + \dfrac{1}{B(a,b)}[R(a,b) - \log 16]\Big\}, \end{cases} \tag{2.14}$$

with equality in each instance if and only if $a = b = 1/2$.

In what follows, we let

$$m_a(r) \equiv \frac{\pi}{2\sin \pi a} r'^2 F(a, 1-a; 1; r^2) F(a, 1-a; 1; r'^2) \tag{2.15}$$

for $a \in (0, 1/2]$ and $r \in (0,1)$. This function is the counterpart of the function

$$m(r) \equiv \frac{2}{\pi} r'^2 \mathcal{K}(r)\mathcal{K}'(r), \quad 0 < r < 1, \tag{2.16}$$

which plays an important role in the study of the distortion functions in quasiconformal theory (cf. [AVV2] and [Hü]). Clearly, $m_{1/2}(r) \equiv m(r)$. As we shall see, the function $m_a(r)$ has applications in the study of $F(a,b;c;x)$, $\mu_a(r)$, and the solution $s = \varphi^a_{1/p}(r)$ (see (1.20)) of the generalized modular equation (1.6). We first obtain the derivative formula for $m_a(r)$.

LEMMA 2.2. *For $a \in (0, 1/2]$, and $r \in (0,1)$,*

$$\begin{aligned} m_a'(r) = -\frac{1}{r} - \frac{\pi r}{\sin \pi a} F(a, 1-a; 1; r'^2)[F(a, 1-a; 1; r^2) \\ - 2a(1-a)F(a, 1-a; 2; r^2)]. \end{aligned} \tag{2.17}$$

In particular, for $r \in (0,1)$,

$$m'(r) = -\frac{1}{r} - \frac{4}{\pi r}\mathcal{K}'(r)[\mathcal{K}(r) - \mathcal{E}(r)] = \frac{1}{r}\left[1 - \frac{4}{\pi}\mathcal{K}(r)\mathcal{E}'(r)\right]. \tag{2.18}$$

Proof. By differentiation, (2.11), and (2.12), we get

$$\begin{aligned} m_a'(r) = -\frac{1}{r} - \frac{\pi r}{\sin \pi a} F(a, 1-a; 1; r'^2)[F(a, 1-a; 1; r^2) \\ -2a(1-a)r'^2 F(a+1, 2-a; 2; r^2)]. \end{aligned} \tag{2.19}$$

Hence, (2.17) follows from [R, Theorem 21, p. 60] and (2.19).

From (1.2), (2.11), and (2.19), we obtain

$$m'(r) = -\frac{1}{r} - \frac{4}{\pi} r\mathcal{K}'(r)\left[\mathcal{K}(r) - \frac{\pi}{4} r'^2 F\left(\frac{3}{2}, \frac{3}{2}; 2; r^2\right)\right]$$
$$= -\frac{1}{r} - \frac{4}{\pi} r\mathcal{K}'(r)\left[\mathcal{K}(r) - \frac{r'^2}{r}\frac{d\mathcal{K}(r)}{dr}\right].$$

Hence, (2.18) follows from the well-known derivative formula for $\mathcal{K}(r)$ and the *Legendre relation* [BF, 110.10]. □

It is well known that the function $f(r) \equiv \mathcal{K}(r)/\log(4/r')$ is strictly decreasing from $[0,1)$ onto $(1, \pi/\log 16]$ (cf. [AVV2, Theorem 3.21(10)] and [QVa, Theorem 1.4]). The next result provides an analogue of this property for $F(a,b;a+b;x)$.

THEOREM 2.3. *Let $a, b \in (0,\infty)$ with $c = a+b$. Then the function*

$$f(x) \equiv F(a,b;c;x)/[R(a,b) - \log(1-x)]$$

is strictly decreasing from $[0,1)$ onto $(1/B(a,b), 1/R(a,b)]$. In particular, $\mathcal{K}(r)/\log(4/r')$ is strictly decreasing from $[0,1)$ onto $(1, \pi/\log 16]$.

Proof. Clearly, $f(0) = 1/R(a,b)$. Since

$$F(a_1, b_1; c_1; 1) = \frac{\Gamma(c_1)\Gamma(c_1 - a_1 - b_1)}{\Gamma(c_1 - a_1)\Gamma(c_1 - b_1)} \tag{2.20}$$

if $c_1 > a_1 + b_1$ and if $c_1 \neq 0, -1, -2, \ldots$ (see [AS, 15.1.20, p. 213] or [R, Theorem 18, p. 49]), by l'Hôpital's rule, [R, Theorem 21, p. 60], and by (2.2), we obtain

$$f(1^-) = \frac{ab}{c}\lim_{x\to 1^-}(1-x)F(a+1, b+1; c+1; x)$$
$$= \frac{ab}{c}F(a,b;c+1;1) = \frac{ab\Gamma(c+1)\Gamma(1)}{c\Gamma(a+1)\Gamma(b+1)} = \frac{1}{B(a,b)}.$$

Next, by differentiation, (2.11), and [R, Theorem 21, p. 60], we get

$$f'(x) = \frac{F(a,b;c+1;x)}{(1-x)[R(a,b) - \log(1-x)]^2} f_1(x), \tag{2.21}$$

where

$$f_1(x) = \frac{ab}{c}[R(a,b) - \log(1-x)] - \frac{F(a,b;c;x)}{F(a,b;c+1;x)}.$$

Clearly, $f_1(0) = [abR(a,b)/c] - 1$, and $(1-x)[R(a,b) - \log(1-x)]^2$ is strictly decreasing in x on $(0,1)$ by (2.7). By (2.20) and (2.2), it holds that

$$F(a,b;c+1;1) = c/[abB(a,b)]. \tag{2.22}$$

By l'Hôpital's rule, we have

$$\lim_{x\to 1^-} \frac{1}{\sqrt{1-x}}\left[\frac{ab}{c}B(a,b)F(a,b;c+1;x) - 1\right] \tag{2.23}$$
$$= -2\frac{(ab)^2 B(a,b)}{c(c+1)}\lim_{x\to 1^-}\sqrt{1-x}F(a+1, b+1; c+2; x) = 0.$$

It follows from (2.8), (2.22), and (2.23) that

$$f_1(1^-) = \lim_{x\to 1^-} \frac{1}{F(a,b;c+1;x)} \left\{ \frac{ab}{c}[R(a,b) - \log(1-x) \right.$$
$$\left. -B(a,b)F(a,b;c;x)]F(a,b;c+1;x) + F(a,b;c;x)\left[\frac{ab}{c}B(a,b)F(a,b;c+1;x) - 1\right]\right\}$$
$$= \frac{1}{F(a,b;c+1;1)} \lim_{x\to 1^-} \left\{ \sqrt{1-x}F(a,b;c;x) \cdot \frac{[abB(a,b)F(a,b;c+1;x)/c] - 1}{\sqrt{1-x}} \right\} = 0.$$

Again we use the formula $F(a,b;a+b+1;z) = (1-z)F(a+1,b+1;a+b+1;z)$ from [R, Theorem 21, p. 60], and because $c = a+b$ we obtain

$$f_1'(x) = \frac{ab}{c(1-x)} - \frac{1}{F(a,b;c+1;x)^2}\left[\frac{ab}{c}F(a+1,b+1;c+1;x)F(a,b;c+1;x)\right.$$
$$\left. - \frac{ab}{c+1}F(a,b;c;x)F(a+1,b+1;c+2;x)\right]$$
$$= \frac{ab}{c+1}\frac{F(a,b;c;x)F(a+1,b+1;c+2;x)}{F(a,b;c+1;x)^2} > 0.$$

Hence f_1 is strictly increasing from $(0,1)$ onto $([abR(a,b)/c] - 1, 0)$, and the monotonicity of f follows from (2.21).

Take $a = b = 1/2$. Then $R(a,b) = \log 16$ by (2.7), $B(a,b) = \pi$ by (2.3), and hence, the second conclusion follows. □

LEMMA 2.4. *Let $a \in (0,1/2]$. Then the function $f(x) \equiv (2-x)F(a,1-a;1;x)^2 - 2$ is strictly increasing and convex from $(0,1)$ onto $(0,\infty)$ if and only if $a = 1/2$. If $a \in (0,1/2)$, then there exists a unique $x_0 \in (0,1)$ such that f is strictly decreasing on $(0,x_0]$, and increasing on $[x_0,1)$, so that f' has a unique zero on $(0,1)$.*

Proof. Clearly, $f(0) = 0$ and $f(1^-) = \infty$. Differentiation gives

$$f'(x) = F(a,1-a;1;x)g(x), \tag{2.24}$$

where

$$g(x) = 2a(1-a)(2-x)F(a+1,2-a;2;x) - F(a,1-a;1;x).$$

Using the series expansion of $F(a,b;c;x)$, we get

$$g(x) = 4\sum_{n=0}^{\infty} \frac{(a,n+1)(1-a,n+1)}{(n+1)!n!}x^n$$
$$- 2\sum_{n=0}^{\infty}\frac{n(a,n)(1-a,n)}{(n!)^2}x^n - \sum_{n=0}^{\infty}\frac{(a,n)(1-a,n)}{(n!)^2}x^n$$
$$= \sum_{n=0}^{\infty}\frac{(a,n)(1-a,n)}{(n+1)!n!}[2n^2+n+4a(1-a)-1]x^n.$$

Since $2n^2+n+4a(1-a)-1 > 0$ for all $n \in \mathbf{N}$, we see that g is strictly increasing on $(0,1)$. Clearly, $g(0) = 4a(1-a)-1 = -(2a-1)^2$. By [R, Theorem 21, p. 60] and (2.22), we find

$$g(1^-) = \lim_{x\to 1^-}\frac{1}{1-x}[2a(1-a)(2-x)F(a,1-a;2;x) - (1-x)F(a,1-a;1;x)] = \infty.$$

Hence, the result follows from (2.24). □

We now prove some properties of the function $m_a(r)$ defined by (2.15).

THEOREM 2.5. *Let $a \in (0, 1/2]$, and $C = R(a)/2$. Then we have the following conclusions:*

1. *The function $f(r) \equiv m_a(r) + \log r$ is strictly decreasing and concave from $(0, 1)$ onto $(0, C)$. In particular, for all $a \in (0, 1/2]$ and $r \in (0, 1)$,*

$$C(1 - r) < m_a(r) + \log r < C. \tag{2.25}$$

2. *The function $g(r) \equiv \{C - [m_a(r) + \log r]\}/r$ is strictly increasing from $(0, 1)$ onto $(0, C)$.*
3. *The function $h(r) \equiv m_a(r)/\log(1/r)$ is strictly increasing from $(0, 1)$ onto $(1, \infty)$.*
4. *For any fixed $t \in (0, 1)$, the function $G(r) \equiv m_a(rt) - m_a(r)$ is strictly increasing from $(0, 1)$ onto $(-\log t, m_a(t))$. In particular, for all $a \in (0, 1/2]$ and $r, t \in (0, 1)$,*

$$\max\{m_a(r) - \log t,\ m_a(t) - \log r\} < m_a(rt) < m_a(r) + m_a(t), \tag{2.26}$$

and

$$m_a(r) - \log r < m_a(r^2) < 2m_a(r). \tag{2.27}$$

Proof. 1. It follows from (2.17) that

$$-\frac{\sin \pi a}{\pi} f'(r) = f_1(r) f_2(r) \equiv f_3(r), \tag{2.28}$$

where

$$f_1(r) = rF(a, 1-a; 1; r'^2) \quad \text{and} \quad f_2(r) = F(a, 1-a; 1; r^2) - 2a(1-a)F(a, 1-a; 2; r^2).$$

Clearly, $f_2(0) = 1 - 2a(1 - a) = a^2 + (1 - a)^2$. By (2.22), we see that $f_2(1^-) = \infty$. Using the series expansion of $F(a, b; c; x)$, we get

$$f_2(r) = \sum_{n=0}^{\infty} \frac{(a, n)(1 - a, n)}{(n + 1)!n!} [n + a^2 + (1 - a)^2] r^{2n},$$

from which it can be easily seen that f_2 is strictly increasing from $(0, 1)$ onto $(a^2 + (1 - a)^2, \infty)$. Hence, by [QVu2, Lemma 2.15(1)], f_3 is a product of two positive and strictly increasing functions on $(0, 1)$ so that the monotonicity and concavity of f follows from (2.28).

The limiting value $f(1^-) = 0$ is clear. It follows from (2.8) and (2.3) that

$$\begin{aligned} f(0^+) &= \frac{1}{2} \lim_{r \to 0^+} [B(a, 1 - a) r'^2 F(a, 1 - a; 1; r^2) F(a, 1 - a; 1; r'^2) + \log r^2] \\ &= \frac{1}{2} \lim_{r \to 0^+} \Bigg\{ [B(a, 1 - a) F(a, 1 - a; 1; r'^2) + \log r^2] \\ &\quad + B(a, 1 - a) r^2 F(a, 1 - a; 1; r'^2) \cdot \frac{r'^2 F(a, 1 - a; 1; r^2) - 1}{r^2} \Bigg\} \\ &= \frac{1}{2} R(a, 1 - a) = \frac{1}{2} R(a) = C, \end{aligned}$$

since by l'Hôpital's rule

$$\lim_{r\to 0^+}\frac{1}{r^2}[r'^2F(a,1-a;1;r^2)-1]=a(1-a)-1.$$

The double inequality (2.25) is clear.

2. Let $g_1(r)=C-[m_a(r)+\log r]=C-f(r)$ and $g_2(r)=r$. Then $g_1(0^+)=g_2(0)=0$, and

$$g_1'(r)/g_2'(r)=-f'(r),$$

which is strictly increasing on $(0,1)$ by part 1. Hence, the monotonicity of g follows from the monotone version of l'Hôpital's rule [AVV2, Theorem 1.24].

Clearly, $g(1^-)=R(a)/2$. By l'Hôpital's rule and (2.28), $g(0^+)=-f'(0^+)=0$.

3. By (2.22), (2.3), [R, Theorem 21, p. 60], and l'Hôpital's rule, we have

$$h(0^+)=\frac{\pi}{2\sin\pi a}\lim_{r\to 0^+}\frac{F(a,1-a;1;r'^2)}{\log(1/r)}=\frac{a(1-a)\pi}{\sin\pi a}F(a,1-a;2;1)=1.$$

Since $\lim_{r\to 1}r'^2/\log(1/r)=2$, we see that $h(1^-)=\infty$.

Next, let $h_1(r)=\log(1/r)$. Then $m_a(1^-)=h_1(1)=0$, and by (2.17),

$$\frac{m_a'(r)}{h_1'(r)}=1+\frac{\pi}{\sin\pi a}rf_1(r)f_2(r)\equiv h_2(r),$$

where f_1 and f_2 are as in (2.28). From the proof of part 1, we see that h_2 is strictly increasing on $(0,1)$, and hence, so is h by [AVV2, Theorem 1.24].

4. Let $x=rt$. Then $x<r$, and

$$rG'(r)=xm_a'(x)-rm_a'(r)=G_1(r)=xf'(x)-rf'(r),$$

where f is as in part 1. From the proof of part 1, we see that $uf'(u)$ is strictly decreasing on $(0,1)$. Hence, $G_1(r)>0$ for all $r\in(0,1)$ so that the monotonicity of G follows.

Clearly, $G(1^-)=m_a(t)$. Since

$$G(r)=[m_a(x)+\log x]-[m_a(r)+\log r]-\log t,$$

it follows from part 1 that $G(0^+)=-\log t$.

The inequalities in (2.26) and (2.27) are clear. □

Remark 2. For $a=1/2$, parts 1–3 of Theorem 2.5 reduce to corresponding well-known results for the function $m(r)$ (cf. [AVV2, Theorem 3.30 parts 1, 2, and 4]).

3. Proofs of the main theorems. In this section, we prove the results stated in section 1.

3.1. Proof of Theorem 1.1. Consider the function

$$f(r)\equiv\mu_a(r)-arth\ r'=\mu_a(r)+\log r-\log(1+r') \tag{3.1}$$

for $a\in(0,1/2]$ and $r\in(0,1)$. In [QVu2, Theorem 1.23(1)], it was shown that f is strictly decreasing and concave from $(0,1)$ onto $(0,[R(a)-\log 4]/2)$.

Let $r_1 = \varphi_2(r')$. Then by [AVV2, Theorem 10.5(4)], $r_1 = 2\sqrt{r'}/(1+r'), r' = \varphi_{1/2}(r_1) = (1-r_1')/(1+r_1')$ and $r = \varphi_2(r_1') = 2\sqrt{r_1'}/(1+r_1')$ so that

$$f(r) = \mu_a\left(\frac{2\sqrt{r_1'}}{1+r_1'}\right) - \frac{1}{2}\log\frac{1+\varphi_{1/2}(r_1)}{1-\varphi_{1/2}(r_1)} = \mu_a\left(\frac{2\sqrt{r_1'}}{1+r_1'}\right) + \frac{1}{2}\log r_1'. \tag{3.2}$$

Let $g(x) = 2\mu_a(2\sqrt{x}/(1+x)) - \mu_a(x)$ for $x \in (0,1)$ and $a \in (0,1/2]$. Then (3.2) can be written as

$$f(r) - \frac{1}{2}\log r_1' - \frac{1}{2}arth\ r_1 = \frac{1}{2}[g(r_1') + f(r_1')],$$

that is

$$f(r) - \frac{1}{2}\log(1+r_1) = \frac{1}{2}[g(r_1') + f(r_1')]. \tag{3.3}$$

Similarly, putting $r_2 = \varphi_2(r_1) = \varphi_4(r')$, we get

$$f(r_1') - \frac{1}{2}\log(1+r_2) = \frac{1}{2}[g(r_2') + f(r_2')],$$

and hence, by (3.3),

$$f(r) - \frac{1}{2}\log(1+r_1) - \frac{1}{4}\log(1+r_2) = \frac{1}{2}g(r_1') + \frac{1}{4}g(r_2') + \frac{1}{4}f(r_2').$$

Generally, assuming

$$f(r) - \sum_{k=1}^{n-1}\frac{1}{2^k}\log(1+r_k) = \sum_{k=1}^{n-1}\frac{1}{2^k}g(r_k') + \frac{1}{2^{n-1}}f(r_{n-1}') \tag{3.4}$$

for $n \in \mathbf{N}$ and $n \geq 2$, we let $r_n = \varphi_2(r_{n-1}) = \varphi_{2^n}(r')$, and from (3.4) it follows that

$$f(r) - \sum_{k=1}^{n}\frac{1}{2^k}\log(1+r_k) = \sum_{k=1}^{n}\frac{1}{2^k}g(r_k') + \frac{1}{2^n}f(r_n'). \tag{3.5}$$

Hence, by induction, (3.5) holds for all $n \in \mathbf{N}$, $a \in (0,1/2]$, and $r \in (0,1)$.

It follows from (3.5), Theorem 2.1, and [QVu2, Theorems 1.14(1) and 1.23(1)] that

$$\begin{aligned}\frac{1}{2^{n+1}}[R(a) - \log 4](1-r_n') < f(r) &- \sum_{k=1}^{n}\frac{1}{2^k}\log(1+r_k)\\ &< \frac{1}{2}[R(a) - \log 16]\sum_{k=1}^{n}\frac{1}{2^k} + \frac{1}{2^n}\left[\frac{1}{2}R(a) - \log 2\right]\\ &= \frac{1}{2}[R(a) - \log 16]\left(1 - \frac{1}{2^n}\right) + \frac{1}{2^n}\left[\frac{1}{2}R(a) - \log 2\right]\\ &= \frac{1}{2}[R(a) - \log 16] + \frac{1}{2^n}\log 4.\end{aligned}$$

Letting $n \to \infty$, we get

$$\sum_{k=1}^{\infty} \frac{1}{2^k} \log(1+r_k) \le f(r) \le \sum_{k=1}^{\infty} \frac{1}{2^k} \log(1+r_k) + \frac{1}{2}[R(a) - \log 16]. \tag{3.6}$$

The double inequality (1.10) now follows from (3.1) and (3.6). Equality (1.11) follows immediately, if we set $a = 1/2$ in (1.10) and apply (2.7).

If $a \in (0, 1/2)$, then by (2.7) and (1.11), the first equality in (1.10) cannot hold since as a function of a, $\mu_a(r)$ is strictly decreasing from $(0, 1/2]$ onto $[\mu(r), \infty)$ by [QVu2, Theorem 1.22]. Suppose that the second equality in (1.10) holds for some $a \in (0, 1/2)$ and for all $r \in (0, 1)$. Then it follows from (1.10) and (1.11) that

$$\mu_a(r) = \mu(r) + \frac{1}{2}[R(a) - \log 16]. \tag{3.7}$$

Letting $r \to 1^-$ in (3.7), we obtain

$$R(a) \equiv \log 16, \quad 0 < a < 1/2.$$

This is a contradiction since $R(a) = \log 16$ if and only if $a = 1/2$ (see the equality case of (2.7)). Consequently, the equality holds in (1.10). □

The next result follows immediately from Theorem 1.1.

COROLLARY 3.1. *For $a \in (0, 1/2]$ and $r \in (0, 1)$,*

$$\mu(r) \le \mu_a(r) \le \mu(r) + \frac{1}{2}(R(a) - \log 16), \tag{3.8}$$

with equality if and only if $a = 1/2$.

3.2. Proof of Theorem 1.2. 1. First we observe that $f(r)$ can be rewritten as

$$f(r) = \frac{m_a(r)}{\log(1/r)} \cdot \frac{1}{r' F(a, 1-a; 1; r^2)^2}. \tag{3.9}$$

Since $r' F(a, 1-a; 1; r^2)$ is strictly decreasing in r from $(0, 1)$ onto $(0, 1)$ by [ABRVV, Theorem 1.7], the right side of (3.9) is a product of two positive and strictly increasing functions on $(0, 1)$ by Theorem 2.5, part 3. Hence, part 1 follows.

2. The monotonicity and the limiting values of g follows from part 1. In order to prove the convexity of g, we let $x = 1/r$ for $r \in (1, \infty)$. Then, by differentiation and (2.10), we obtain

$$\begin{aligned} g'(r) = -g_1(x) &\equiv \frac{x}{[\log(1/x)]^2} \left\{ \frac{\log(1/x)}{[x' F(a, 1-a; 1; x^2)]^2} - \mu_a(x) \right\} \\ &= -\frac{x}{\log(1/x)} \cdot \frac{1}{[x' F(a, 1-a; 1; x^2)]^2} \cdot \left[\frac{m_a(x)}{\log(1/x)} - 1 \right]. \end{aligned} \tag{3.10}$$

Clearly, $x/\log(1/x)$ is strictly increasing in x on $(0, 1)$. Since $x' F(a, 1-a; 1; x^2)$ is strictly decreasing in x on $(0, 1)$ by [ABRVV, Theorem 1.7], it follows from Theorem 2.5, part 3 that $g_1(x)$ is a product of three positive and strictly increasing functions of x on $(0, 1)$. Hence, g_1 is strictly decreasing in r on $(1, \infty)$ so that g' is strictly increasing in r on $(1, \infty)$, and the convexity of g follows.

Next, using (3.10), we get

$$G'(r) = -g'(1/r)/r^2 = g_1(r)/r^2 \tag{3.11}$$
$$= \frac{1}{[r'F(a,1-a;1;r^2)]^2} \cdot \frac{m_a(r)+\log r}{r[\log(1/r)]^2}$$
$$= \frac{1}{r\log(1/r)} \cdot \frac{1}{[r'F(a,1-a;1;r^2)]^2} \cdot \left[\frac{m_a(r)}{\log(1/r)} - 1\right].$$

By l'Hôpital's rule, Theorem 2.5, part 1, and (2.28),

$$\lim_{r\to 0^+} \frac{[m_a(r)+\log r]/r}{[\log(1/r)]^2} = \frac{1}{2}\lim_{r\to 0^+} \frac{[\pi r f_3(r)/\sin\pi a] + m_a(r) + \log r}{r\log(1/r)} = \infty,$$

where f_3 is as in (2.28). Hence, from the third equality in (3.11) we see that $G'(0^+) = \infty$. On the other hand, it follows from the fourth equality in (3.11) and Theorem 2.5, part 3 that $G'(1^-) = \infty$. Consequently, $G'(r)$ is neither decreasing nor increasing on $(0,1)$, and the assertion about $G(r)$ follows.

3. By differentiation and (2.10), we obtain

$$-h'(r) = h_1(r) \equiv \frac{1}{\{[(R(a)/2)+\log(1/r)]r'F(a,1-a;1;r^2)\}^2} \cdot \frac{[R(a)/2]-[m_a(r)+\log r]}{r}.$$

Since $r'F(a,1-a;1;r^2)$ is strictly decreasing from $(0,1)$ onto $(0,1)$ by [ABRVV, Theorem 1.7], by Theorem 2.5, part 2, $h_1(r)$ is a product of two positive and strictly increasing functions on $(0,1)$. Hence, the monotonicity and concavity of h follow.

Clearly, $h(1^-) = 0$, while the limit $h(0^+) = 1$ follows from part 1.

4. The function $H(r)$ can be written as

$$H(r) = \frac{\pi}{\sin\pi a} \frac{1}{\sqrt{r'}F(a,1-a;1;r^2)} \cdot \frac{F(a,1-a;1;r'^2)}{R(a)-\log r^2}. \tag{3.12}$$

Since $\sqrt{r'}F(a,1-a;1;r^2) = \sqrt[4]{1-r^2}F(a,1-a;1;r^2)$ is strictly decreasing from $(0,1)$ onto $(0,1)$ [ABRVV, Theorem 1.7], the result for H follows from (3.12), Theorem 2.3, and (2.3).

The first and second inequalities in (1.12) follow from the monotonicity of H and part 3, respectively. □

3.3. Proof of Theorem 1.3. 1. Put $x = e^{-r}$. Then $x \in (0,1)$, and by differentiation and (2.10), we get

$$f'(r) = [x'F(a,1-a;1;x^2)]^{-2},$$

which is strictly decreasing in r on $(0,\infty)$ by [ABRVV, Theorem 1.7]. Hence, the concavity of f follows.

The inequalities (1.13), (1.14) and their equality case are clear.

2. This follows immediately from part 1.

3. Let $x = rt/(1+r't')$. Then $x < rt < r$ and

$$\frac{dx}{dr} = \frac{t(r'+t')}{r'(1+r't')^2} = \frac{xx'}{rr'}.$$

Differentiation and (2.10) give

$$h'(r) = \frac{x'F(a,1-a;1;x^2)^2 - r'F(a,1-a;1;r^2)^2}{rx'[r'F(a,1-a;1;r^2)F(a,1-a;1;x^2)]^2},$$

which is positive for all $r \in (0,1)$ since $r'F(a,1-a;1;r^2)^2$ is strictly decreasing on $(0,1)$ by [ABRVV, Theorem 1.7] and since $x < r$. Hence the monotonicity of h follows.

Clearly, $h(1^-) = \mu_a(t)$. It follows from [QVu2, Corollary 3.12] that

$$h(0^+) = \lim_{r\to 0^+} \left\{ [\mu_a(x) + \log x] - [\mu_a(r) + \log r] + \log \frac{1+r't'}{t} \right\} = \log \frac{1+t'}{t} = arth\, t'.$$

The second and third inequalities in (1.16) are clear, while the first inequality in (1.16) follows from [QVu2, Theorem 1.23(1)].

4. Let $x = \sqrt{r}$. Then by differentiation, (1.5), (2.10), and (2.15), we get

$$\frac{r}{2}\left[\frac{\pi}{\sin \pi a} r'x'F(a,1-a;1;r^2)F(a,1-a;1;x'^2)\right]^2 G'(r) = m_a(r) - 2m_a(x),$$

which is negative for all $r \in (0,1)$ by (2.27). This yields the monotonicity of G.

By (1.5) and (2.8), we have

$$G(0^+) = \lim_{r\to 0^+} \frac{F(a,1-a;1;r'^2)}{F(a,1-a;1;1-r)} = \lim_{r\to 0^+} \frac{\log r^2}{\log r} = 2,$$

and

$$G(1^-) = \lim_{r\to 1^-} \frac{F(a,1-a;1;r)}{F(a,1-a;1;r^2)} = \lim_{r\to 1^-} \frac{\log(1-r)}{\log(1-r^2)} = 1.$$

The double inequality (1.17) is clear. □

3.4. Proof of Theorem 1.4. 1. Let $f_1(r) = r[r'F(a,1-a;1;r^2)]^2$. Then by differentiation and [R, Theorem 21, p. 60],

$$[F(a,1-a;1;r^2)]^{-1} f_1'(r) = f_3(r) \equiv r'^2 F(a,1-a;1;r^2) - 2rf_2(r),$$

where f_2 is as in the proof of Theorem 2.5, part 1. In the proof of Theorem 2.5, part 1, we have shown that f_2 is strictly increasing from $(0,1)$ onto $(a^2 + (1-a)^2, \infty)$. Hence, by [ABRVV, Theorem 1.7], f_3 is strictly decreasing from $(0,1)$ onto $(-\infty, 1)$, so that there exists a unique $r_0 \in (0,1)$ such that f_1 is strictly increasing on $(0, r_0]$, and decreasing on $[r_0, 1)$ with range $(0, f_1(r_0)]$.

Next, by (2.10),

$$f'(r) = -1/f_1(r), \tag{3.13}$$

which is strictly increasing on $(0, r_0]$, and decreasing on $[r_0, 1)$. This yields the assertion about f.

Let $f_4(r) = r[r'F(a,1-a;1;r'^2)]^2$, and

$$f_5(r) = r'^2 F(a,1-a;1;r^2) + 2a(1-a)r^2 F(a,1-a;2;r^2).$$

Then by differentiation, (2.10), and [R, Theorem 21, p. 60], we get

$$\frac{d}{dr}\left(\frac{1}{f(r)}\right) = \left(\frac{2\sin \pi a}{\pi}\right)^2 \frac{1}{f_4(r)}, \tag{3.14}$$

and

$$[F(a,1-a;1;r'^2)]^{-1}f_4'(r)=f_6(r)\equiv r'^2F(a,1-a;1;r'^2)-2f_5(r'). \tag{3.15}$$

Using the series expansion of $F(a,b;c;x)$, we obtain

$$\begin{aligned}f_5(r)&=(1-r^2)\sum_{n=0}^{\infty}\frac{(a,n)(1-a,n)}{(n!)^2}r^{2n}+2a(1-a)\sum_{n=0}^{\infty}\frac{(a,n)(1-a,n)}{(n+1)!n!}r^{2(n+1)}\\&=\sum_{n=0}^{\infty}\frac{(a,n)(1-a,n)}{(n!)^2}r^{2n}-\sum_{n=1}^{\infty}\frac{(a,n-1)(1-a,n-1)}{[(n-1)!]^2}r^{2n}\\&\quad+2a(1-a)\sum_{n=1}^{\infty}\frac{(a,n-1)(1-a,n-1)n}{(n!)^2}r^{2n}\\&=1+\sum_{n=1}^{\infty}\frac{(a,n-1)(1-a,n-1)}{(n!)^2}\{a(1-a)-[a^2+(1-a)^2]n\}r^{2n},\end{aligned}$$

which is strictly decreasing on $(0,1)$ since

$$a(1-a)-[a^2+(1-a)^2]n\le 3a(1-a)-1<0$$

for all $n\in\mathbf{N}$. Hence, f_6 is strictly decreasing on $(0,1)$. Clearly, $f_6(1^-)=-2$. By (2.22) and (2.3), $f_5(1^-)=(2\sin\pi a)/\pi$ so that $f_6(0^+)=\infty$. Thus, there exists a unique $r_1\in(0,1)$ such that f_4 is strictly increasing on $(0,r_1]$ and decreasing on $[r_1,1)$ with $f_4(0^+)=f_4(1^-)=0$. Consequently, the assertion about $1/f(r)$ follows from (3.14).

2. Differentiation and (2.10) give

$$g'(r)=\frac{F(a,1-a;1;r^2)+1}{[r'F(a,1-a;1;r^2)]^2}\cdot\frac{F(a,1-a;1;r^2)-1}{r},$$

which is a product of two positive and strictly increasing functions on $(0,1)$ since by (1.1)

$$\frac{1}{r}[F(a,1-a;1;r^2)-1]=\sum_{n=1}^{\infty}\frac{(a,n)(1-a,n)}{(n!)^2}r^{2n-1},$$

and since $r'F(a,1-a;1;r^2)$ is strictly decreasing on $(0,1)$ by [ABRVV, Theorem 1.7]. Hence, the monotonicity and convexity of g follow.

The limiting values follow from [QVu2, Corollary 3.12].

3. By differentiation and (2.10), we get

$$-h'(r)=\Big[\frac{r}{r'}\exp(\mu_a(r))\Big]\cdot\frac{1}{r'F(a,1-a;1;r^2)^2}\cdot\frac{1-[r'F(a,1-a;1;r^2)]^2}{r},$$

which is a product of three positive and strictly increasing functions on $(0,1)$ by part 2 and [QVu2, Lemma 2.16 (2) & (3)]. Hence the monotonicity and concavity of h follow.

Clearly, $h(1^-)=1$. By part 2, we have $h(0^+)=\exp(R(a)/2)$.

4. Differentiation gives

$$G'(r)=\frac{(2-r^2)F(a,1-a;1;r^2)^2-2}{2r[r'F(a,1-a;1;r^2)]^2}, \tag{3.16}$$

which is positive for all $r \in (0,1)$ if and only if $a = 1/2$, by Lemma 2.4. Hence, part 4i and the piecewise monotonicity stated in part 4iii of Theorem 1.4 follow from (3.16) and Lemma 2.4.

Next, let

$$G_1(r) = [(2-r)F(a,1-a;1;r)^2 - 2]/r, \quad G_2(r) = (2-r)F(a,1-a;1;r)^2 - 2,$$

and $G_3(r) = r$. Then $G_2(0) = G_3(0) = 0$, and by (2.24),

$$G_2'(r)/G_3'(r) = F(a,1-a;1;r)G_4(r), \tag{3.17}$$

where

$$G_4(r) = 2a(1-a)(2-r)F(a+1,2-a;2;r) - F(a,1-a;1;r).$$

It was proved in the proof of Lemma 2.4 that G_4 is strictly increasing from $(0,1)$ onto $(-(2a-1)^2, \infty)$. Hence, if $a = 1/2$, then by (3.17), $G_2'(r)/G_3'(r)$ is strictly increasing on $(0,1)$, and so is G_1 by [AVV2, Theorem 1.24] with $G_1(0^+) = G_4(0) = 0$. Consequently, by (3.16) and [ABRVV, Theorem 1.7], G' is strictly increasing on $(0,1)$ if $a = 1/2$, so that G is convex on $(0,1)$.

Now suppose that $a \in (0, 1/2)$. By (3.16), (2.24), [R, Theorem 21, p. 60], and by differentiation, we get

$$\begin{aligned} 2r'^4 F(a,1-a;1;r^2)^3 G''(r) = G_5(r) \equiv 2[r'F(a,1-a;1;r^2)]^2 G_4(r^2) \\ - G_1(r^2)[(1-3r^2)F(a,1-a;1;r^2) + 4a(1-a)r^2 F(a,1-a;2;r^2)]. \end{aligned} \tag{3.18}$$

Clearly, $G_5(1^-) = \infty$. It follows from (3.17), l'Hôpital's rule, and the discussion in the previous paragraph that

$$G_5(0^+) = 2G_4(0) - G_1(0^+) = 2G_4(0) - G_4(0) = -(2a-1)^2 < 0.$$

Therefore, by (3.18), G' is not monotone on $(0,1)$ if $a \in (0,1/2)$, so that part 4ii and the second assertion in part 4iii follow.

Finally, the limiting values of G follow from [QVu2, Corollary 3.12]. □

Conjecture 1. For $a \in (0,1/2)$, the function $g(r) \equiv \mu_a(r)/arth\ \sqrt[4]{r'}$ is strictly increasing from $(0,1)$ onto $(1,\infty)$.

4. Open problem. It was proved in [QVa, Theorem 1.4(1)&(2)] that the function $f(r) \equiv \mathcal{K}(r)/\log(4/r')$ is not only strictly decreasing but also strictly concave on $[0,1)$, while $g(r) \equiv f(r')$ is strictly convex on $(0,1]$. In Theorem 2.3, we obtained the analogue of the monotonicity property of $f(r)$ for its generalization $h(r) \equiv F(a,b;a+b;r^2)/[R(a) - \log(1-r^2)]$ for $a,b \in (0,\infty)$. What are the analogues of the concavity of $f(r)$ and the convexity of $g(r)$ for $h(r)$ and $h(r')$, respectively? Our computational work seems to show that for small values of a and b, say $a+b<1$, $h(r)$ and $h(r')$ are concave and convex on $(0,1)$, respectively, and for large values of a and b, say $a,b \in (1,\infty)$, $h(r)$ and $h(r')$ are convex and concave on $(0,1)$, respectively, while for some values of a and b, both $h(r)$ and $h(r')$ are neither concave nor convex on $(0,1)$.

Acknowledgments. This research was completed during the first author's visit to the University of Helsinki on grants from the Finnish Academy of Sciences and Letters, the Academy of Finland, and from the Commission on Development and Exchanges of International Mathematical Union. He wishes to express his thanks to

these institutions and to the Department of Mathematics of the University of Helsinki for its excellent research facilities. Both authors are indebted to the referees, whose remarks were very valuable.

REFERENCES

[AS] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Dover, New York, 1965.

[ABRVV] G. D. Anderson, R. W. Barnard, K. C. Richards, M. K. Vamanamurthy, and M. Vuorinen, *Inequalities for zero-balanced hypergeometric functions*, Trans. Amer. Math. Soc., 347 (1995), pp. 1713–1723.

[AQV] G. D. Anderson, S.-L. Qiu and M. K. Vamanamurthy, *Elliptic integral inequalities, with applications*, Constr. Approx., 14 (1998), pp. 195–207.

[AQVV] G. D. Anderson, S.-L. Qiu, M. K. Vamanamurthy, and M. Vuorinen, *Generalized elliptic integrals and modular equations*, Pacific J. Math., (1999), to appear.

[AV] G. D. Anderson and M. K. Vamanamurthy, *Some properties of quasiconformal distortion functions*, New Zealand J. Math., 24 (1995), pp. 1–15.

[AVV1] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Hypergeometric functions and elliptic integrals*, in Current Topics in Analytic Function Theory, H. M. Srivastava and S. Owa, eds., World Scientific Publ. Co., Singapore, London, 1992, pp. 48–85.

[AVV2] G. D. Anderson, M. K. Vamanamurthy, and M. Vuorinen, *Conformal Invariants, Inequalities, and Quasiconformal Maps*, John Wiley & Sons, New York, 1997.

[Ao] K. Aomoto, *Hypergeometric functions—the past, today, and... (from the complex analytic point of view)*, Sugaku Expositions, 9 (1996), pp. 99–116.

[Ask1] R. Askey, *Handbooks of Special Functions, A Century of Mathematics in America*, Part III, P. Duren, ed., AMS, Providence, RI, 1989, pp. 369–391.

[Ask2] R. Askey, *Ramanujan and hypergeometric and basic hypergeometric series*, Ramanujan International Symposium on Analysis, N. K. Thakare, ed., Pune, India, 1987, pp. 1–83; Russian Math. Surveys, 451 (1990), pp. 37–86.

[BPV] R. Balasubramanian, S. Ponnusamy, and M. Vuorinen, *Functional inequalities for the quotients of hypergeometric functions*, J. Math. Anal. Appl., 218 (1998), pp. 256–268.

[Be1] B. C. Berndt, *Ramanujan's Notebooks*, Vol. I, Springer-Verlag, Berlin, Heidelberg, New York, 1985.

[Be2] B. C. Berndt, *Ramanujan's Notebooks*, Vol. II, Springer-Verlag, Berlin, Heidelberg, New York, 1989.

[Be3] B. C. Berndt, *Ramanujan's Notebooks*, Vol. III, Springer-Verlag, Berlin, Heidelberg, New York, 1991.

[Be4] B. C. Berndt, *Ramanujan's Notebooks*, Vol. IV, Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[Be5] B. C. Berndt, *Ramanujan's theory of theta-functions*, Centre de Recherches Mathématiques, CRM Proc. Lecture Notes, 1 (1993), pp. 1–63.

[BBG] B. C. Berndt, S. Bhargava, and F. G. Garvan, *Ramanujan's theories of elliptic functions to alternative bases*, Trans. Amer. Math. Soc., 347 (1995), pp. 4163–4244.

[BB] J. M. Borwein and P. B. Borwein, *Pi and the AGM*, John Wiley & Sons, New York, 1987.

[Bo] F. Bowman, *Introduction to Elliptic Functions with Applications*, Dover, New York, 1961.

[BF] P. F. Byrd and M. D. Friedman, *Handbook of Elliptic Integrals for Engineers and Physicists*, 2nd ed., Grundlehren Math. Wiss. 67, Springer-Verlag, Berlin, Göttingen, Heidelberg, New York, 1971.

[CC] D. V. Chudnovsky and G. V. Chudnovsky, *Hypergeometric and modular function identities, and new rational approximations to a continued fraction expansions of classical constants and functions*, in A Tribute to Emil Grosswald—Number Theory and Related Analysis, M. Knopp and M. Sheingorn, eds., Contemp. Math. 143, AMS, Providence, RI, 1993, pp. 117–162.

[E] R. J. Evans, *Ramanujan's second notebook: Asymptotic expansions for hypergeometric series and related functions*, in Ramanujan Revisited, Proceedings of the Ramanujan Centenary Conference at the University of Illinois at Urbana–Champaign, G.

E. Andrews, R. A. Askey, B. C. Berndt, R. G. Ramanathan, and R. A. Rankin, eds., Academic Press, Boston, 1988, pp. 537–560.

[Ga] F. G. Garvan, *Ramanujan's theories of elliptic functions to alternative bases–a symbolic excursion*, J. Symbolic Comput., 20 (1995), pp. 517–536.

[HP] J. Hersch and A. Pfluger, *Généralisation du lemme de Schwarz et du principe de la measure harmonique pour les fonctions pseudo-analytique*, C. R. Acad. Sci. Paris, 234 (1952), pp. 43–45.

[Hü] O. Hübner, *Remarks on a paper of Lawrynowicz on quasiconformal mappings*, Bull. de L'Acad. Polon. des Sci., 18 (1970), pp. 183–186.

[J] C. G. J. Jacobi, *Fundamenta Nova Theoriae Functionum Ellipticarum* (1829), Gesammelte Werke, Berlin, 1881.

[LV] O. Lehto and K. I. Virtanen, *Quasiconformal Mappings in the Plane*, 2nd ed., Grundlehren Math. Wiss. 126, Springer-Verlag, New York, Berlin, 1973.

[M] G. J. Martin, *The distortion theorem for quasiconformal mappings, Schottky's theorem and holomorphic motions*, Proc. Amer. Math. Soc., 125 (1997), pp. 1095–1103.

[QVa] S.-L. Qiu and M. K. Vamanamurthy, *Elliptic integrals and the modulus of Grötzsch ring*, Panamer. Math. J., 5 (1995), pp. 41–60.

[QVu1] S.-L. Qiu and M. Vuorinen, *Landen inequalities for hypergeometric functions*, Nagoya Math. J., 154 (1999), pp. 31–56.

[QVu2] S.-L. Qiu and M. Vuorinen, *Duplication inequalities for the ratios of hypergeometric functions*, Forum Math., 11 (1999), pp. 1–25.

[R] E. D. Rainville, *Special Functions*, Chelsea Publ. Co., New York, 1960.

[VV] M. K. Vamanamurthy and M. Vuorinen,*Functional inequalities, Jacobi products, and quasiconformal maps*, Illinois J. Math., 38 (1994), pp. 394–419.

[Var] V. S. Varadarajan, *Linear meromorphic differential equations: A modern point of view*, Bull. Amer. Math. Soc., 33 (1996), pp. 1–42.

[Va] A. Varchenko, *Multidimensional hypergeometric functions and their appearance in conformal field theory, algebraic K-theory, algebraic geometry, etc.*, in Proceedings of the International Congr. Math., Kyoto, Japan, 1990, pp. 281–300.

[Vu] M. Vuorinen, *Conformal Geometry and Quasiregular Mappings*, Lecture Notes in Math. 1319, Springer–Verlag, Berlin, New York, 1988.

[WW] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4th ed., Cambridge University Press, London, l958.

[WZ] H. S. Wilf and D. Zeilberger, *An algorithmic proof theory for hypergeometric (ordinary and "q") multisum/integral identities*, Invent. Math., 108 (1992), pp. 575–633.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1076–1114

UNSTABLE OSCILLATORY-TAIL WAVES IN COLLISIONLESS PLASMAS*

YAN GUO† AND WALTER A. STRAUSS‡

Abstract. Consider a collisionless relativistic neutral plasma. An oscillatory-tail equilibrium is a state whose magnetic field connects two different constant states at $x = -\infty$ and $x = +\infty$ and whose electric field oscillates as $x \to -\infty$. We prove that such a state is nonlinearly dynamically unstable under certain perturbations of the initial data.

Key words. collisionless plasma, instability, BGK mode, oscillatory electric field, relativistic particles

AMS subject classifications. 35L60, 82C21, 82D10

PII. S0036131098333918

1. Introduction.

A collisionless plasma of electrons and ions is described by the Vlasov–Maxwell system. In such a plasma, collisions are relatively rare; here we assume no collisions at all. In many plasmas, some of the particles are expected to travel at relativistic speeds. However, in a nonrelativistic Vlasov model, particles can travel at arbitrarily great speed. We avoid this anomaly by assuming a relativistic model. Thus we consider the relativistic Vlasov–Maxwell system (RVM):

$$
\begin{aligned}
&\partial_t f_\pm + \hat{v}_\pm \cdot \nabla_x f_\pm + e_\pm(\mathbf{E} + \hat{v}_\pm \times \mathbf{B}) \cdot \nabla_v f_\pm = 0,\\
&\partial_t \mathbf{E} - c\ \mathrm{curl}\ \mathbf{B} = -\mathbf{j} = -\int_{\mathbf{R}^3} [e_+ \hat{v}_+ f_+ + e_- \hat{v}_- f_-] dv,\\
&\partial_t \mathbf{B} + c\ \mathrm{curl}\ \mathbf{E} = 0,\\
&\mathrm{div}\ \mathbf{E} = \rho = \int_{\mathbf{R}^3} [e_+ f_+ + e_- f_-] dv, \quad \mathrm{div}\ \mathbf{B} = 0,
\end{aligned}
\tag{1}
$$

where $m_\pm$ and $e_\pm$ are the masses and charges of the ions $(+)$ and electrons $(-)$, respectively. Here f_+ is the distribution of the ions, f_- the distribution of the electrons at time t, $\mathbf{E}$ the electric field, $\mathbf{B}$ the magnetic field, $\mathbf{x}$ the position, $\mathbf{v}$ the momentum, and $\hat{\mathbf{v}}_\pm$ the velocity.

For notational simplicity, we set all constants equal to 1 and $e_\pm = \pm 1$, so that the velocity is $\hat{\mathbf{v}} = \mathbf{v}/\sqrt{1 + |\mathbf{v}|^2}$. We consider the simplest scenario where there can be a magnetic field, the so-called $1\frac{1}{2}$-dimensional system $1\frac{1}{2}$RVM, where the position is $\mathbf{x} = (x, 0, 0)$, the momentum is $\mathbf{v} = (v_1, v_2, 0)$, the electric field is $\mathbf{E} = (E_1, E_2, 0)$, and the magnetic field is $\mathbf{B} = (0, 0, B)$.

A fundamental feature of this collisionless model is the multiplicity of its steady states. The question of their dynamical stability has played a crucial role in plasma physics and is related to plasma control and turbulence. We consider equilibria of the

*Received by the editors February 9, 1998; accepted for publication (in revised form) September 16, 1998; published electronically August 26, 1999.
http://www.siam.org/journals/sima/30-5/33391.html

†Division of Applied Mathematics and LCDS, Brown University, Providence, RI 02912 (guoy@cfm.brown.edu), and Department of Mathematics, Princeton University, Princeton, NJ 08544. The research of this author was supported in part by NSF grant 96-23253 and an NSF Postdoctoral Fellowship.

‡Department of Mathematics and LCDS, Brown University, Providence, RI 02912 (wstrauss@math.brown.edu). The research of this author was supported in part by NSF grant 97-03695.

form

$$f_\pm = \mu_\pm(\langle v\rangle \mp \Phi(x), v_2 \pm \Psi(x)),\ E_1 = \partial_x\Phi,\quad E_2 = 0,\quad B = \partial_x\Psi, \tag{2}$$

which are a generalization of the Bernstein–Greene–Kruskal (BGK) modes [BGK]. It was observed in [GR] that such a state satisfies $1\frac{1}{2}$RVM if Φ and Ψ satisfy the coupled pair of ODEs

$$\begin{aligned}\Phi_{xx} &= \int_{\mathbf{R}^2}[\mu_+(\langle v\rangle - \Phi, v_2 + \Psi) - \mu_-(\langle v\rangle + \Phi, v_2 - \Psi)]dv,\\ \Psi_{xx} &= -\int_{\mathbf{R}^2}\hat{v}_2[\mu_+(\langle v\rangle - \Phi, v_2 + \Psi) - \mu_-(\langle v\rangle + \Phi, v_2 - \Psi)]dv.\end{aligned} \tag{3}$$

Furthermore some of these states are of the oscillatory-tail type, that is, $\Phi(x)$ approaches a periodic solution $\beta(x)$ as $x \to -\infty$ and a constant as $x \to +\infty$, while Ψ approaches two different constants at $+\infty$ and $-\infty$. The periodic function $\beta(x)$ satisfies the following ODE:

$$\beta_{xx} = \rho = \int_{\mathbf{R}^2}[\mu_+(\langle v\rangle - \beta, v_2) - \mu_-(\langle v\rangle + \beta, v_2)]dv_1dv_2. \tag{4}$$

We consider quintuples $u = [f_+, f_-, E_1, E_2, B]$. On such quintuples we consider the sum of the L^1 norms of the five components. Let $\beta(x)$ be a periodic solution to (4) so that

$$\Gamma_0 = [\mu_\pm(\langle v\rangle \mp \beta(x), v_2); \partial_x\beta, 0; 0] \tag{5}$$

is a periodic solution to (3), where $E_1 = \partial_x\beta$, $E_2 \equiv 0$, $B = 0$. Assume a solution $\Phi(x), \Psi(x)$ to (3) so that

$$\Gamma \equiv \Gamma(\Phi, \Psi)(x, v) = [\mu_\pm(\langle v\rangle \mp \Phi(x), v_2 \pm \Psi(x)); \partial_x\Phi, 0; \partial_x\Psi], \tag{6}$$

where $E_1 = \partial_x\Phi, E_2 \equiv 0, B = \partial_x\Psi$, satisfies

$$\lim_{x\to-\infty}\{|\Phi - \beta| + |\partial_x(\Phi - \beta)| + |\Psi| + |\partial_x\Psi|\} = 0. \tag{7}$$

Thus Γ_0 is a periodic solution, while Γ has an oscillatory tail. As mentioned above, such oscillatory-tail solutions exist [GR].

Our goal is to prove the following theorem on the instability of such states Γ. We will consider the class of solutions of (1) given by Theorem 4 of the appendix.

THEOREM 1. *Let $\mu_\pm$ satisfy $\mu_\pm = O(\langle v\rangle^{-l})$ for some $l > 3$ and* (11), (12), (14), (61), *and* (62). *Let $\beta(x)$ in* (5) *have period P_β and $\|\beta\|_{C^2}$ be sufficiently small. If Γ satisfies* (7), *then there exist $\epsilon_0 > 0$ and $C_1 > 0$ and a family of solutions $u^\delta(t) = [f_+^\delta, f_-^\delta, E_1^\delta, E_2^\delta, B^\delta]$ of $1\frac{1}{2}$RVM for $0 < \delta < \delta_0$ (with $f_\pm^\delta \geq 0$), as well as a family of intervals K^δ, each of length $2P_\beta$, such that*

$$||u^\delta(0) - \Gamma||_{W^{1,1}(\mathbf{R}\times\mathbf{R}^2)} < \delta$$

but

$$\sup_{0\leq t\leq C_1|\ln\delta|}\|u^\delta(t) - \Gamma\|_{L^1(K^\delta\times\mathbf{R}^2)} \geq \epsilon_0.$$

Notice that the solution escapes from a δ neighborhood of the equilibrium in a time $O(|\log \delta|)$. This property characterizes an exponential instability. Conditions (11), (61), and (62) require that $\mu_\pm$ be smooth and positive and satisfy certain decay conditions at infinity.

A major difficulty of this problem is that the spatial variable is unbounded. Therefore the growing plane wave solutions do not belong to any natural function space like L^p and correspond to the *continuous spectrum* of the linearized operator. Now the asymptotic solution Γ_0 satisfies the reduced system with $E_2 = B = 0$ and periodic boundary conditions, which system we denote by $1\frac{1}{4}$RVM. This system is analyzed by linearization to produce a linear system which we denote by $1\frac{1}{4}$L. While the periodic linear problem still has continuous spectrum (because of the unbounded v variable), it has less of it and we prove that it has some unstable point spectrum. We reduce the original problem to the periodic one using the causality property of RVM.

Thus there are four levels of instability that appear in this paper. The simplest level is that of the system linearized around a simple homogeneous state. This level is easily reduced to a dispersion relation for which explicit sufficient (and almost necessary) conditions are found following [P]. The second level is that of the system $1\frac{1}{4}$L linearized around the periodic equilibrium $\beta(x)$. The third level is that of the nonlinear system $1\frac{1}{4}$RVM. Our goal is to reach the fourth level, the full nonlinear system $1\frac{1}{2}$RVM which possesses the oscillatory-tail equilibrium solutions. Normally it is expected that exponential growth of the linearized system (the second level) should imply the nonlinear instability of the equilibrium (the third level). We are aware of no previous work other than ours [GS1], [GS2], [GS3], [GS4] that proves the instability of spatially dependent equilibria either on the linearized or the nonlinear level.

Although this paper is closely related to our previous ones, there are some major differences. Our earlier papers treated only the one-dimensional case, and there were no magnetic effects. In this paper there is a second momentum variable v_2 as well as a magnetic field. Furthermore the asymptotic behavior of the equilibrium as $x \to -\infty$ now is oscillatory instead of constant.

Some other related references are the following. Several classes of equilibria are constructed in [GR]. Stability of a homogeneous equilibrium μ within the Vlasov–Poisson theory has been discussed since the 1960s, beginning with the linear instability analysis of [P]. The nonlinear stability has been proven only for monotonically decreasing distributions μ that do not depend on the space variable. [G1] and [G2] prove the stability of various spatially dependent states. In particular, [G1] proves the stability of flat-tail equilibria, which possess nonoscillatory behavior as $x \to \pm\infty$. [G2] proves, for the three-dimensional RVM, the stability of general axially symmetric magnetic equilibria that are given variationally as minimizers of a natural action functional.

Section 2 of this paper is devoted to the periodic equilibria Γ_0 of the periodic system $1\frac{1}{4}$RVM. In section 3 such an equilibrium is proven to be linearly unstable due to a point eigenvalue. The analysis is similar to that of [GS3] except that the variable v_2 appears nontrivially in many places. The density profiles $\mu_\pm$ are required to decay in v_2 as well as in the energy $\langle v \rangle$. Section 4 is devoted to the regularity and certain pointwise estimates of the unstable eigenfunctions. In section 5 we prove the nonlinear instability of the periodic equilibrium Γ_0. Again, the variable v_2 appears in a significant way in some key places, for instance in the crucial Lemma 15. In section 6 we finally prove the main theorem by deriving the nonlinear instability of the oscillatory-tail equilibrium Γ in the $1\frac{1}{2}$ dimensional system from the nonlinear

instability of Γ_0 within the periodic $1\frac{1}{4}$ dimensional system. The key property that is required is that $\Gamma \to \Gamma_0$ as $x \to -\infty$.

2. BGK periodic waves. We consider the system that is obtained from $1\frac{1}{2}$ RVM by letting $E_2 \equiv 0, B \equiv 0$ and by considering functions that are even in v_2. This is the system $1\frac{1}{4}$ RVM:

$$
\begin{aligned}
&(\partial_t + \hat{v}_1\partial_x \pm E_1\partial_{v_1})f_\pm = 0,\\
&\partial_t E_1 = -j_1 = -\int\int \hat{v}_1(f_+ - f_-)dv_1dv_2,\\
&\partial_x E_1 = \rho = \int\int (f_+ - f_-)dv_1dv_2,
\end{aligned}
\tag{8}
$$

where

$$f(t,x,v_1,v_2) = f(t,x,v_1,-v_2), \tag{9}$$

$$f(t,x+P,v_1,v_2) = f(t,x,v_1,v_2) \tag{10}$$

with $P > 0$ fixed. We emphasize that $\langle v\rangle = (1+v_1^2+v_2^2)^{1/2}$ and $\hat{v}_1 = v_1\langle v\rangle^{-1}$ in (8). This system is well posed. (See appendix.)

We assume $\mu_\pm$ are C^2 functions on $\mathbf{R}^2$ such that

$$\sum_{|\sigma|\le 2} |\partial^\sigma \mu_\pm(s,v_2)| \le Cs^{-\gamma}\langle v_2\rangle^{-\tilde{\gamma}}, \qquad \gamma > 1, \gamma+\tilde{\gamma} > 2, C > 0, \tag{11}$$

$$\int_{\mathbf{R}^2} [\mu_+(\langle v\rangle, v_2) - \mu_-(\langle v\rangle, v_2)]dv = 0, \qquad \mu_\pm(\cdot, v_2) = \mu_\pm(\cdot, -v_2) \ge 0. \tag{12}$$

Note that the decay condition (11) implies that

$$\mu_\pm(\langle v\rangle, v_2) \in W^{2,1}(\mathbf{R}^2).$$

We let the potential function $H(\phi)$ satisfy

$$-H'(\phi) = \int_{\mathbf{R}^2} [\mu_+(\langle v\rangle - \phi, v_2) - \mu_-(\langle v\rangle + \phi, v_2)]dv$$

and we define $H(0) = 0$. This integral is easily shown to be finite.

Remark. With slightly more decay assumed on $\mu_\pm$, we derive the following formula (13) for H. Substituting $s = \langle v\rangle \mp \phi$ and letting $\langle v_2\rangle = \sqrt{1+v_2^2}$, we obtain

$$
\begin{aligned}
-H'(\phi) =& \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [\mu_+(\langle v\rangle - \phi, v_2) - \mu_-(\langle v\rangle + \phi, v_2)]dv_1dv_2\\
=& 2\int_{-\infty}^{\infty}\int_{\langle v_2\rangle - \phi}^{\infty} \mu_+(s,v_2)\frac{(s+\phi)dsdv_2}{\sqrt{(s+\phi)^2 - 1 - v_2^2}}\\
&-2\int_{-\infty}^{\infty}\int_{\langle v_2\rangle + \phi}^{\infty} \mu_-(s,v_2)\frac{(s-\phi)dsdv_2}{\sqrt{(s-\phi)^2 - 1 - v_2^2}}\\
=& 2\frac{\partial}{\partial\phi}\int_{-\infty}^{\infty}\int_{\langle v_2\rangle - \phi}^{\infty} \mu_+(s,v_2)\sqrt{(s+\phi)^2 - 1 - v_2^2}dsdv_2\\
&+2\frac{\partial}{\partial\phi}\int_{-\infty}^{\infty}\int_{\langle v_2\rangle + \phi}^{\infty} \mu_-(s,v_2)\sqrt{(s-\phi)^2 - 1 - v_2^2}dsdv_2.
\end{aligned}
$$

Therefore,

$$H(\phi) = C - 2\int_{-\infty}^{\infty}\int_{\langle v_2\rangle-\phi}^{\infty}\mu_+(s,v_2)\sqrt{(s+\phi)^2-1-v_2^2}\,ds dv_2$$
$$-2\int_{-\infty}^{\infty}\int_{\langle v_2\rangle+\phi}^{\infty}\mu_-(s,v_2)\sqrt{(s-\phi)^2-1-v_2^2}\,ds dv_2, \tag{13}$$

where $C = 2\int_{-\infty}^{\infty}\int_{\langle v_2\rangle}^{\infty}[\mu_+(s,v_2)+\mu_-(s,v_2)]\sqrt{s^2-1-v_2^2}\,ds dv_2$. Notice that as $\phi \to \pm\infty$, $H(\phi) \to -\infty$ unless $\mu_\pm \equiv 0$.

We also have

$$H''(0) = \int_{\mathbf{R}^2}[\partial_e\mu_+(\langle v\rangle, v_2) + \partial_e\mu_-(\langle v\rangle, v_2)]dv,$$

where e represents the first argument of $\mu_\pm$. (In the proof given below, we show this is finite.)

LEMMA 1 (periodic BGK equilibria). *Let $\mu_\pm$ satisfy* (11) *and* (12), *and let*

$$\left(\frac{2\pi}{P_0}\right)^2 = \int_{\mathbf{R}^2}[\partial_e\mu_+(\langle v\rangle, v_2) + \partial_e\mu_-(\langle v\rangle, v_2)]dv > 0. \tag{14}$$

Then there exists $\delta_0 > 0$ such that for all $\delta < \delta_0$, there exists a periodic function $\beta(x)$ with period P_β satisfying (4), *and*

$$|\beta|_\infty = \delta, \qquad \lim_{\delta\to 0} P_\beta = P_0, \tag{15}$$
$$\beta(0) = \beta(P_\beta) = \min_{0\le x\le P_\beta}\beta(x), \quad \beta(P_\beta/2) = \max_{0\le x\le P_\beta}\beta(x).$$

Here P_0 is defined by (14) *and we can take $\delta_0 = \sup\{s : H''(s) > 0\}$.*

For fixed β, we shall sometimes drop the subscript on P_β.

Proof. Clearly $H(\cdot)$ is a C^3 function, since

$$\int_{\mathbf{R}^2}\langle v\rangle^{-\gamma}\langle v_2\rangle^{-\tilde\gamma}dv_1 dv_2 < \infty.$$

Consider the ODE (4). Obviously $\beta \equiv 0$ is a solution to (4) by (12). Note that $H'(0) = 0$ and

$$H''(0) = \int_{\mathbf{R}^2}[\partial_e\mu_+(\langle v\rangle, v_2) + \partial_e\mu_-(\langle v\rangle, v_2)]dv > 0$$

by (14). Thus the origin is a center for the ODE. In the phase space (β, β_x), let $(a,0)$ and $(b,0)$ be two points that lie on a periodic orbit such that $a < 0$ and $b > 0$. Then we have

$$\frac{1}{2}P_\beta = \int_0^b \frac{du}{(2H(b)-2H(u))^{1/2}} + \int_a^0 \frac{du}{(2H(a)-2H(u))^{1/2}}.$$

In order to prove (15), we shall take the limits as $b \to 0$ and $a \to 0$. We have already proven $H'(0) = 0$ and $H''(0) = (2\pi/P_0)^2$. Expanding in Taylor series around 0, we have for $|u| \le b$,

$$H(b) - H(u) = H''(0)(b^2-u^2)/2 + O(b^3).$$

Hence

$$\int_0^b \{2H(b)-2H(u)\}^{-1/2}du = H''(0)^{-1/2}\int_0^b (b^2-u^2)^{-1/2}du + O(b)$$
$$= (P_0/2\pi)(\pi/2) + O(b) = P_0/4 + O(b).$$

We can handle the second integral similarly. Hence $P_\beta/2 = P_0/2 + O(|b|+|a|)$. We finally arrange β to have its minimum at the ends of $[0, P_\beta]$ and its maximum at the middle by a translation of x. This proves (15). □

The linearized form of $1\frac{1}{4}$ RVM (see (8)) around the homogeneous state $[\mu_\pm(\langle v\rangle, v_2)$, $E_1 \equiv 0]$ is

$$\begin{aligned}
&(\partial_t + \hat{v}_1\partial_x)g_\pm = \mp E_1\partial_{v_1}\mu_\pm(\langle v\rangle, v_2),\\
&\partial_t E_1 = -\int_{\mathbf{R}^2} \hat{v}_1(g_+ - g_-)dv = -j_1,\\
&\partial_x E_1 = \int_{\mathbf{R}^2}(g_+ - g_-)dv = \rho.
\end{aligned} \tag{16}$$

We emphasize that system (16) is *not* RVM in (1) linearized around the BGK equilibrium but only around the corresponding *homogeneous* state.

LEMMA 2 (homogeneous growing modes). *Let $\mu_\pm(e, v_2)$ satisfy conditions* (11), (12), *and* (14). *Then there exists a growing exponential solution for* (16) *of period* $2P_0$:

$$g_\pm(t) = \pm\frac{\hat{v}_1\partial_e\mu_\pm(\langle v\rangle, v_2)}{\hat{v}_1 - \omega_0/k}e^{ikx - i\omega_0 t}, \quad E_1(t) = -ike^{ikx-i\omega_0 t}, \tag{17}$$
$$\int_0^{2P_0} E_1(t,x)dx = \int_0^{2P_0} j_1(t,x)dx = 0.$$

Here $k = \frac{\pi}{P_0} > 0$ and ω_0 is a pure imaginary number. Moreover,

$$\operatorname{Im}\omega_0 > 0. \tag{18}$$

Proof. Notice that the function

$$Z(i\lambda) = \int_{\mathbf{R}^2} \frac{\hat{v}_1[\partial_e\mu_+(\langle v\rangle, v_2) + \partial_e\mu_-(\langle v\rangle, v_2)]dv}{\hat{v}_1 - i\lambda}$$

is real and continuous for $0 \le \lambda < \infty$ by integration by parts because $\partial_e\mu_\pm$ are odd functions of v_1. Moreover, $Z(0) = (2\pi/P_0)^2$ and $\lim_{\lambda\to\infty} Z(i\lambda) = 0$. Hence there exists $\lambda > 0$ such that $Z(i\lambda) = (\pi/P_0)^2$. It follows directly that the following triple is a solution of (16):

$$g_\pm = \pm\frac{\hat{v}_1\partial_e\mu_\pm(\langle v\rangle, v_2)}{\hat{v}_1 - i\lambda}e^{k[ix+\lambda t]}, \quad E_1 = -ike^{k[ix+\lambda t]}.$$

Here $k = \pi/P_0$. We deduce the lemma by letting $\omega_0 = ik\lambda, k \ne 0$. Clearly this is a growing mode since it has the factor $\exp(k\lambda t)$ with $k\lambda > 0$. □

Remark. If $Z(i\lambda_0) \ge (2\pi/P_0)^2$ for some $\lambda_0 > 0$, then there is a growing mode with period P_0 instead of $2P_0$.

3. Linear instability for periodic BGK waves. In this section, we shall prove the instability for the linearized Vlasov–Maxwell system around periodic BGK waves by using a perturbation method. We formulate the linearized problem equivalently in terms of the Poisson equation and a complicated operator $\mathcal{C}$ involving the Vlasov characteristics. Then through detailed estimates along the trajectories, we conclude that the linear operator is a nice perturbation of the homogeneous case, whereby it indeed has a growing mode.

Let $\beta = \beta(x)$ be any given periodic BGK wave with period P. We study the linearized Vlasov–Maxwell system around the generalized BGK wave

$$[\mu_\pm(\langle v\rangle \mp \beta, v_2), \beta_x, 0, 0] \equiv [f_\pm, E_1, E_2, B].$$

In this section, we assume $E_2 = B = 0$, and we denote E_1 by E. Thus we have the system ($1\frac{1}{4}$L):

$$\begin{aligned} &(\partial_t + \hat{v}_1\partial_x \pm \beta'\partial_{v_1})g_\pm \pm E\partial_{v_1}\mu_\pm(\langle v\rangle \mp \beta, v_2) = 0,\\ &\partial_t E = -\int_{\mathbf{R}^2} \hat{v}_1[g_+ - g_-]dv_1dv_2 = -j_1,\\ &\partial_x E = \int_{\mathbf{R}^2} (g_+ - g_-)dv_1dv_2 = \rho, \end{aligned} \tag{19}$$

with the P-periodic boundary condition. We will consider pairs of functions $g = [g_+(x,v), g_-(x,v)]$ and triples $u = [g_+(x,v), g_-(x,v), E(x)]$. We sometimes write $e = \langle v\rangle \mp \beta$ and $v = (v_1, v_2)$.

DEFINITION. *Let $\mathcal{M}$ be the Banach space of triples $u = [g_+(x,v), g_-(x,v), E(x)]$ of finite measures on $\mathbf{R}_P \times \mathbf{R}^2$, $\mathbf{R}_P \times \mathbf{R}^2$, and $\mathbf{R}_P$, respectively, which are periodic in x with period P, respectively, and satisfy*

$$\begin{aligned} &\int_0^P \int_{\mathbf{R}^2} g_- dvdx = \int_0^P \int_{\mathbf{R}^2} g_+ dvdx \quad \text{(neutrality)},\\ &\partial_x E = \int_{\mathbf{R}^2} [g_+ - g_-]dv \quad \text{(Poisson equation)},\\ &g_\pm(x, v_1, v_2) = g_\pm(x, v_1, -v_2) \quad \text{(evenness)}. \end{aligned} \tag{20}$$

We denote the norm $\|u\|_m = \|g_+\|_m + \|g_-\|_m + |E|_m$ where $\|\cdot\|_m$ and $|\cdot|_m$ are the corresponding measure norms in $\mathbf{R}_P \times \mathbf{R}^2$ and $\mathbf{R}_P$.

DEFINITION. *We define the operator $\mathcal{A}$ acting on pairs $g = [g_+, g_-]$ into pairs and the operator $\mathcal{K}$ acting on E into pairs by*

$$\begin{aligned} \mathcal{A}g &= \begin{pmatrix} \mathcal{A}_+(g_+) \\ \mathcal{A}_-(g_-) \end{pmatrix} = \begin{pmatrix} \hat{v}_1\partial_x g_+ + \beta'\partial_{v_1} g_+ \\ \hat{v}_1\partial_x g_- - \beta'\partial_{v_1} g_- \end{pmatrix},\\ \mathcal{K}E &= \begin{pmatrix} \mathcal{K}_+E \\ \mathcal{K}_-E \end{pmatrix} = \begin{pmatrix} \partial_{v_1}\mu_+(\langle v\rangle - \beta, v_2)E \\ -\partial_{v_1}\mu_-(\langle v\rangle + \beta, v_2)E \end{pmatrix}. \end{aligned} \tag{21}$$

Here the domain of $\mathcal{A}$ is the set of pairs of measures g such that $\mathcal{A}g$ is a pair of measures. Furthermore, we define $\mathcal{L}$ from triples to triples by

$$\mathcal{L}u = \begin{pmatrix} \mathcal{A}_+g_+ + \mathcal{K}_+E \\ \mathcal{A}_-g_- + \mathcal{K}_-E \\ \int_{\mathbf{R}^2} \hat{v}_1[g_+ - g_-]dv \end{pmatrix}. \tag{22}$$

LEMMA 3 (linearized well-posedness). *Let $\mu_\pm$ satisfy* (11), (12) *and let β be any solution of* (4) *of period P. If $u_0 \in \mathcal{M}$, there is unique solution $u(t) \in \mathcal{M}$ of*

$$\frac{du}{dt} + \mathcal{L}u = 0, \qquad u(0) = u_0.$$

Sketch of the proof. We split the operator $\mathcal{L}$ as

$$\mathcal{L}u = \begin{pmatrix} \mathcal{A}(g) \\ \int_{\mathbf{R}^2} \hat{v}_1[g_+ - g_-]dv \end{pmatrix} + \begin{pmatrix} \mathcal{K}(E) \\ 0 \end{pmatrix} \equiv \mathcal{L}_1 u + \mathcal{L}_2 u. \tag{23}$$

Notice that the Vlasov operator $e^{-\mathcal{A}t}$ has norm 1 and $|j_1|_m \le \|g_+\|_m + \|g_-\|_m$. The operator $\mathcal{L}_1$ thus generates a strongly continuous semigroup on $\mathcal{M}$ with

$$\|e^{-\mathcal{L}_1 t}u_0\|_m \le C(1+t)\|u_0\|_m. \tag{24}$$

Now

$$|\partial_x E|_m = \|\rho\|_m \le \|g_+\|_m + \|g_-\|_m,$$

so that $\mathcal{L}_2$ is a compact operator on $\mathcal{M}$ and our lemma thus follows.

We introduce the characteristics $X^\pm(t;0,x',v_1',v_2')$ and $V^\pm(t;0,x',v_1',v_2')$ as the solutions of

$$\frac{dX^\pm}{dt} = \hat{V}^\pm, \quad \frac{dV_1^\pm}{dt} = \pm\beta'(X^\pm), \quad \frac{dV_2^\pm}{dt} = 0,$$
$$X^\pm(0) = x', \quad V_1^\pm(0) = v_1', \quad V_2^\pm(0) = v_2'.$$

Let $L^1(\mathbf{R}_P)$ be the space of P-periodic integrable functions of x and let $L^1(\mathbf{R}_P \times \mathbf{R}^2)$ be the similar space of functions of x and $v = (v_1, v_2)$ with the norms

$$|h(\cdot)|_1 = \int_0^P |h(x)|dx, \qquad \|h(\cdot,\cdot)\|_1 = \int_0^P \int_{\mathbf{R}^2} |h(x,v)|dvdx.$$

Let $W^{1,1}(\mathbf{R}_P)$ and $W^{1,1}(\mathbf{R}_P \times \mathbf{R}^2)$ be the subspaces of $L^1(\mathbf{R}_P)$ and $L^1(\mathbf{R}_P \times \mathbf{R}^2)$ with the norms $|h|_{1,1} = |\partial_x h|_1 + |h|_1$ and $\|h\|_{1,1} = \|\partial_x h\|_1 + \|\partial_v h\|_1 + \|h\|_1$.

DEFINITION. *For* $\operatorname{Im}\omega > 0$, *we define*

$$R_\pm = -\int_0^\infty e^{-s\mathcal{A}_\pm} e^{i\omega s}\mathcal{K}_\pm E ds, \tag{25}$$

$$\rho(x) = \int_{\mathbf{R}^2} [R_+(x,v) - R_-(x,v)]dv, \tag{26}$$

$$j_1(x) = \int_{\mathbf{R}^2} \hat{v}_1[R_+(x,v) - R_-(x,v)]dv, \tag{27}$$

$$[\mathcal{C}(\omega,\beta)E](x) = \int_0^x \rho(y)dy + \frac{1}{P}\int_0^P \left\{ \frac{1}{i\omega} j_1(y)dy - \int_0^y \rho(z)dz \right\} dy. \tag{28}$$

LEMMA 4. *Assume* (11) *and* (12).

(a) *If* $\operatorname{Im}\omega > 0$, *then $\mathcal{C}(\omega,\beta)$ is a bounded linear operator on $L^1(\mathbf{R}_P)$.*

(b) *Suppose that $E \in L^1(\mathbf{R}_P)$ satisfies the equation*

$$E = \mathcal{C}(\omega,\beta)E$$

for some $\operatorname{Im}\omega > 0$. *Then there exist* $R_\pm(x,v) \in L^1(\mathbf{R}_P \times \mathbf{R}^2)$ *such that*

$$\begin{aligned} &\hat{v}_1 \partial_x R_\pm \pm \beta \partial_{v_1} R_\pm \pm \partial_{v_1}\mu_\pm(\langle v\rangle \mp \beta, v_2)E(x) = i\omega R_\pm, \\ &j_1 = i\omega E, \quad \partial_x E = \rho. \end{aligned} \tag{29}$$

That is, $-i\omega$ *is an eigenvalue with a positive real part of the linearized Vlasov–Maxwell generator* $-\mathcal{L}$.

(c) *In terms of the characteristics, we have*

$$\int_0^x \rho(y)dy = \int_{\mathbf{R}} K(x,x')E(x')dx', \tag{30}$$

where $K = K^+ + K^-$,

$$K^\pm = -\int_0^\infty \int_{\mathbf{R}^2} \mathcal{H}\partial_{v_1}\mu_\pm(e, v_2')e^{is\omega}dv'ds,$$

where $\mathcal{H} = H(x - X^\pm(s;0,x',v')) - H(-X^\pm(s;0,x',v'))$ *and* $H(\cdot)$ *is the Heaviside function.*

(d) *We also have*

$$\begin{aligned} &\int_0^x j_1(y)dy = \int_{\mathbf{R}^2} J(x,x')E(x')dx', \qquad J = J^+ + J^-, \\ &J^\pm = -\int_0^\infty \int_{\mathbf{R}^2} \hat{V}_1^\pm(s;0,x,v)\mathcal{H}\partial_{v_1}\mu_\pm(e', v_2')e^{is\omega}dv'ds, \end{aligned}$$

where $e' = \langle v'\rangle \mp \beta(x')$.

Remark. Notice that for fixed $(s;0,x',v')$ and x, we have

$$\mathcal{H} = \int_0^x \delta(y - X^\pm(s;0,x',v'))dy,$$

where δ is the Dirac measure. Therefore, $\mathcal{C}(\omega,\beta)$ could also be defined equivalently by

$$\begin{aligned} &\rho(x) = \partial_x[\mathcal{C}(\omega,\beta)E](x) = \int_{\mathbf{R}} E(x')k(x,x')dx', \\ &\text{where } k = k^+ + k^- \text{ and } k^\pm(x,x') = \partial_x K^\pm(x,x'). \end{aligned} \tag{31}$$

Everywhere that $k(x,x')$ is used in the following proofs, it will appear under an integral sign, and therefore all the integrals are rigorous classical expressions.

Proof. Notice that $\mathcal{A}_\pm$ are unbounded linear operators on $L^1(\mathbf{R}_P \times \mathbf{R}^2)$ that generate groups of isometries $e^{-s\mathcal{A}_\pm}$ on $L^1(\mathbf{R}_P \times \mathbf{R}^2)$. So by (25) we have

$$\begin{aligned} \|R_\pm\|_1 &\le \int_0^\infty |e^{i\omega s}|\|E\partial_{v_1}\mu_\pm\|_1 ds \\ &\le (\operatorname{Im}\omega)^{-1}|E|_1 \sup_x \int_{\mathbf{R}^2} |\partial_e \mu_\pm(\langle v\rangle \mp \beta(x), v_2)\hat{v}_1|dv \\ &\le 2(\operatorname{Im}\omega)^{-1}|E|_1 \int_{\mathbf{R}^2} |\partial_e\mu_\pm(e,v_2)|dedv_2 < \infty \end{aligned} \tag{32}$$

for any $E \in L^1(\mathbf{R}_P)$. It follows easily that $\rho, j_1 \in L^1(\mathbf{R}_P)$ and $\mathcal{C}(\omega,\beta)E \in L^1(\mathbf{R}_P)$.

To prove (b), notice that $i\omega$ belongs to the resolvent set of $\mathcal{A}_\pm$, so that $R_\pm$ can be written as

$$R_\pm = \pm(i\omega I - \mathcal{A}_\pm)^{-1}(E\partial_{v_1}\mu_\pm).$$

That is, $R_\pm$ satisfies $\mathcal{A}_\pm R_\pm \pm E\partial_{v_1}\mu_\pm = i\omega R_\pm$, which is the first pair of equations in (29). Integrating (29) over x and v, we have $\int_0^P \int R_\pm dvdx = 0$. Finally, from the assumption $E = \mathcal{C}E$, we obtain

$$\partial_x E = \partial_x \mathcal{C}E = \rho = \int (R_+ - R_-)dv.$$

Notice that by integrating (29) over v, we have $\partial_x j_1 = i\omega\rho = i\omega\partial_x E$. From the definition of $\mathcal{C}$, $i\omega E$ has the same average over P as j_1, so that $i\omega E = j_1$.

We now prove part (c). By (25),

$$\int_{\mathbf{R}^2} R_\pm(x,v)dv = \mp \int_{\mathbf{R}^2}\int_0^\infty e^{-s\mathcal{A}_\pm} E\partial_{v_1}\mu_\pm e^{i\omega s} dsdv.$$

For any function $f(x,v)$,

$$\begin{aligned}
&\int_0^x \int_{\mathbf{R}^2} (e^{-s\mathcal{A}_\pm} f)(y,v)dvdy \\
&= \int_0^x \int_{\mathbf{R}^2} f(X^\pm(0;s,y,v), V^\pm(0;s,y,v))dvdy \\
&= \int_{\mathbf{R}} \int_{\mathbf{R}^2} \mathbf{1}_{0\le y\le x} f(X^\pm(0;s,y,v), V^\pm(0;s,y,v))dvdy \\
&= \int_{\mathbf{R}^2}\int_{\mathbf{R}} [H(x - X^\pm(s;0,x',v')) - H(-X^\pm(s;0,x',v'))]f(x',v')dx'dv'
\end{aligned}$$

by making the change of variables

$$x' = X^\pm(0;s,y,v), \quad v' = V^\pm(0;s,y,v).$$

The inverse transformation is

$$y = X^\pm(s;0,x',v'), \quad v = V^\pm(s;0,x',v')$$

and the Jacobian is 1. Using this identity, we have

$$\begin{aligned}
&\int_0^x \int_{\mathbf{R}^2} R_\pm(y,v)dvdy \\
&= \mp \int_{\mathbf{R}} \int_0^\infty \int_{\mathbf{R}^2} [H(x - X^\pm(s;0,x',v')) - H(-X^\pm(s;0,x',v'))]E(x')\partial_{v_1}\mu_\pm e^{is\omega} dv'dsdx'.
\end{aligned}$$

This immediately implies (c).

To prove (d), we have the similar

$$\int_0^x \int_{\mathbf{R}^2} \hat{v}_1 (e^{-s\mathcal{A}_\pm} f)(y,v)dvdy = \int_{\mathbf{R}^2}\int_{\mathbf{R}} \hat{V}_1^\pm(s;0,x',v')\mathcal{H}f(x',v')dx'dv'$$

so that

$$\int_0^x \int_{\mathbf{R}} \hat{v}_1 R_\pm(y,v)dvdy = \mp \int_{\mathbf{R}}\int_0^\infty \int_{\mathbf{R}^2} \hat{V}_1^\pm(s;0,x',v')\mathcal{H}E(x')\partial_{v_1}\mu_\pm e^{is\omega} dv'dsdx',$$

which immediately implies (d). □

In order to analyze the operator $\mathcal{C}$, we have to estimate along the trajectories. Consider the particle paths given by

$$\frac{dx}{dt} = \hat{v}_1, \qquad \frac{dv_1}{dt} = \pm\beta'(x), \qquad \frac{dv_2}{dt} = 0, \tag{33}$$

whose solutions are $X^{\pm}(t;0,x',v')$, $V_1^{\pm}(t;0,x',v')$, v_2'. We define the *untrapped* region of the + flows as

$$F^+ = \{(x',v')|\ \langle v'\rangle - \beta(x') > 1 - \min\ \beta = a\}.$$

In F^+ the trajectories can go from $-\infty$ to $+\infty$. We also define the *trapped* region of the + flows as

$$T^+ = \{(x',v')|\ \langle v'\rangle - \beta(x') \le 1 - \min\ \beta\},$$

where the flows will never move out of each interval $[nP,(n+1)P]$, by our choice of β in (15). Similarly, we define the *untrapped* region of the − flows as

$$F^- = \{(x',v')|\ \langle v'\rangle + \beta(x') > 1 - \max\ \beta = b\},$$

where the flows can go from $-\infty$ to $+\infty$ and the *trapped* region of the − flows as

$$T^- = \{(x',v')|\ \langle v'\rangle + \beta(x') \le 1 + \max\ \beta\},$$

where the flows will never move out of each interval $[nP - P/2, nP + P/2]$, by our choice of β in (15). Let

$$\Sigma^{\pm}(t,x,x',v_2) = \{v_1' \in \mathbf{R} \mid X^{\pm}(t;0,x',v_1',v_2') = x\}$$

be the initial velocity of a particle travelling from x' to x in time t. Notice that $\Sigma^{\pm}$ could, inside the trapped region $T^{\pm}$, consist of more than one point. The flows with different initial velocities could come back to the same position in the same time, as long as the consumed time interval is a common multiple of their different periods. However, in the untrapped region $\Sigma^{\pm}$ is a single point.

LEMMA 5. (a) *If* $(x',v') \in F^{\pm}$ *and* $v_1' \in \Sigma^{\pm}(t,x,x',v_2')$, *then* $\Sigma^{\pm}(t,x,x',v_2')$ *consists of a unique point* $\mathcal{V}_{1\pm}(t,x,x',v_2')$. *Moreover*

$$\mathcal{V}_{1\pm}(t,x+P,x'+P,v_2') = \mathcal{V}_{1\pm}(t,x,x',v_2'). \tag{34}$$

(b) $\mathcal{C}(\omega,\beta)$ *maps* P*-periodic functions to* P*-periodic functions.*

Notation. We define $\mathcal{V}_{\pm} = (\mathcal{V}_{1\pm}, v_2')$ and

$$\langle \mathcal{V}_{\pm}\rangle = \sqrt{1 + \mathcal{V}_{1\pm}^2 + |v_2'|^2}.$$

We also define the free velocity and the free energy by

$$\hat{\mathcal{V}}_{10}(t,x,x') = (x-x')/t, \tag{35}$$

$$\langle \mathcal{V}_0(t,x,x',v_2')\rangle = \sqrt{1+|v_2'|^2}\left[1 - \left(\frac{x-x'}{t}\right)^2\right]^{-1/2}. \tag{36}$$

Proof. For part (a), without loss of generality we may consider just the + part, since similar arguments apply to the − part. If $v_1' \in \Sigma^+(t, x, x', v_2')$, then

$$\langle V^+\rangle - \beta(X^+) = \langle v'\rangle - \beta(x'), \tag{37}$$

where $X^+ = X^+(0; t, x', v')$, $V^+ = V^+(0; t, x', v')$. Notice that for every $v = (v_1, v_2) \in \mathbf{R}^2$,

$$\langle v\rangle^2 - v_2^2 - 1 = v_1^2, \quad \hat{v}_1 = \pm\frac{\sqrt{\langle v\rangle^2 - 1 - v_2^2}}{\langle v\rangle}.$$

Therefore from the characteristic ODEs (33) and from (37), we have

$$\frac{dX^+}{dt} = \hat{V}_1^+ = \pm\frac{[(\langle v'\rangle - \beta(x') + \beta(X^+))^2 - 1 - v_2^2]^{1/2}}{\langle v'\rangle - \beta(x') + \beta(X^+)}.$$

Because $v_1' \in \Sigma^+(t, x, x', v_2)$, we have

$$t = \pm\int_{x'}^{x} \frac{\langle v'\rangle - \beta(x') + \beta(y)}{[(\langle v'\rangle - \beta(x') + \beta(y))^2 - 1 - v_2^2]^{1/2}}\,dy. \tag{38}$$

Here we let $x = X^+(t; 0, x', v')$ and use + if $t > 0$ and $x > x'$. If we take $x > x'$, then $t > 0$ and $(x', v') \in F^+$. Then $v_1' > 0$. As we choose the plus sign in (38), t is a strictly decreasing function of $\langle v'\rangle$ and a strictly increasing function of x. Hence v_1' is uniquely determined and we write $v_1' = \mathcal{V}_{1+}(t, x, x', v_2)$.

To show the periodicity (34) of $\mathcal{V}_\pm$, notice that $X^+(t; 0, x'+P, v') - P$ satisfies the same ODE and initial conditions as $X^+(t; 0, x', v')$, and hence they are equal. Thus

$$X^+(t; 0, x' + P, v') = X^+(t; 0, x', v') + P. \tag{39}$$

So by the definition of $\mathcal{V}_+$,

$$\mathcal{V}_+(t; x, x', v_2) = \mathcal{V}_+(t; x + P, x' + P, v_2).$$

Similarly for the − case.

Next we shall prove part (b). By (25), it suffices to show that if E has period P, then

$$\int_x^{x+P} dy \int_{\mathbf{R}} dx' E(x')k(y, x') = 0 \tag{40}$$

for all x. Notice that, by the definitions of k and K, by the periodicity of β, and by (39),

$$k^\pm(y, x') = k^\pm(y + P, x' + P).$$

Hence the x-derivatives of both integrals $\int_x^{x+P} dy \int_{\mathbf{R}} dx' E(x')k^\pm(y, x')$ are zero. We thus have for all x,

$$\int_x^{x+P} dy \int_{\mathbf{R}} dx' E(x')k^\pm(y, x') = \int_0^P dy \int_{\mathbf{R}} dx' E(x')k^\pm(y, x').$$

Therefore, in order to prove (40), it suffices to show

$$\int_0^P dx \int_{\mathbf{R}} dx' E(x')k(x, x') = 0. \tag{41}$$

Notice that $[0,P]$ is a period for two connected trapped regions in the + case. If $v_1' \in \Sigma^+(\tau,x,x',v_2)$ and $(x',v') \in F^+$, then $v_1' = \mathcal{V}_{1+}(\tau,x,x',v_2)$ is also an increasing function of x. Hence for $x' \le 0$ and for $P \le x'$, we have

$$\{v_1' |\ 0 \le X^+(\tau;0,x',v_1',v_2) \le P\} = \{\mathcal{V}_+(\tau;x,x',v_2)|\ 0 \le x \le P\}$$
$$= \text{the interval } [\mathcal{V}_+(\tau,0,x',v_2), \mathcal{V}_+(\tau,P,x',v_2)] = J.$$

Now let $0 < x' < P$. For $0 < x' < P$, there is an interval I of velocities v_1' that are trapped. Of course for any $v_1' \in I$,

$$0 < X^+(t;0,x',v') < P$$

for *all* t. Notice that $I \subset J$. In this case, J is the interval from $\mathcal{V}_{1+}(\tau;0,x',v_2) < 0$ to $\mathcal{V}_{1+}(\tau;P,x',v_2) > 0$. Notice that for fixed (τ,x',v_2),

$$\int_0^P \delta(x - X^+(\tau;0,x',v'))dx = \mathbf{1}_J(v_1'),$$

where $\mathbf{1}$ is the standard characteristic function. In order to prove (41), we use the definitions of $k(x,x')$ and $K(x,x')$ below (30). We first treat the + part in (41) as

$$\begin{aligned}
&-\int_0^P dx \int_{\mathbf{R}} dx' E(x') k^+(x,x') \\
&= \int_0^\infty d\tau e^{i\omega\tau} \int_{\mathbf{R}^3} dv' dx' E(x') \partial_{v_1'} \mu_+ \left[\int_0^P dx\ \delta(x - X^+(\tau;0,x',v')) \right] \\
&= \int_{\mathbf{R}} dx' E(x') \int_0^\infty d\tau e^{i\omega\tau} \int_{\mathbf{R}} dv_2 \int_J dv_1' \partial_{v_1'} \mu_+(\langle v' \rangle - \beta(x'), v_2) \\
&= \int_{\mathbf{R}} dx' E(x') \int_0^\infty d\tau e^{i\omega\tau} \int_{\mathbf{R}} dv_2 [\mu_+(\langle \mathcal{V}_+(\tau,P,x',v_2) \rangle - \beta(x'), v_2) \\
&\qquad\qquad -\mu_+(\langle \mathcal{V}_+(\tau,0,x',v_2) \rangle - \beta(x'), v_2)] \\
&= I_1 + I_2
\end{aligned}$$

Letting $x' = z + P$ in I_1, we obtain

$$\begin{aligned}
I_1 &= \int_{\mathbf{R}} dz E(z+P) \int_0^\infty d\tau e^{i\omega\tau} \int_{\mathbf{R}} dv_2 \mu_+(\langle \mathcal{V}_+(\tau,P,z+P,v_2) \rangle - \beta(z+P), v_2) \\
&= \int_{\mathbf{R}} dz E(z) \int_0^\infty d\tau e^{i\omega\tau} \int_{\mathbf{R}} dv_2 \mu_+(\langle \mathcal{V}_+(\tau,0,z,v_2) \rangle - \beta(z), v_2) = -I_2,
\end{aligned}$$

where we have used the P-periodicity of E and β as well as (34). Therefore we conclude that the + part of (41) is zero.

The − part is also zero by the same argument. The only difference is that now we consider the period $[P/2, 3P/2]$ instead of $[0,P]$. This is a period for two connected trapped regions of the − flow. This completes the proof of the lemma. □

Let $X^\pm(\tau;0,x',v')$ be the trajectories in (33) and $X^0(\tau;0,x',v') = x' + \tau \hat{v}_1'$, $V^0(\tau;0,x',v') = v'$ be the unperturbed trajectories (straight lines). Recalling the definitions of $k^\pm$, $\hat{\mathcal{V}}_{10}$ and $\langle \mathcal{V}_0 \rangle$ in (35) and (36), we also define

$$\begin{aligned}
(42) \qquad k_0^\pm(x,x') &= \mp \int_0^\infty \left[\int \delta(x - x' - \hat{v}_1'\tau) \partial_{v_1'} \mu_\pm(\langle v' \rangle, v_2') dv' \right] e^{i\omega\tau} d\tau \\
&= \mp \int_0^\infty \int_{\mathbf{R}} \partial_e \mu_\pm(\langle \mathcal{V}_0 \rangle, v_2) \hat{\mathcal{V}}_{10} \langle \mathcal{V}_0 \rangle^3 (1+v_2^2)^{-1} dv_2 \tau^{-1} e^{i\omega\tau} d\tau.
\end{aligned}$$

Recalling the definitions (21), (25), (26), and (27), we similarly define

$$\mathcal{A}^0 = \hat{v}_1 \partial_x, \qquad \mathcal{K}^0 E = \begin{pmatrix} \mathcal{K}^0_+ E \\ \mathcal{K}^0_- E \end{pmatrix} = \begin{pmatrix} \partial_{v_1} \mu_+(\langle v \rangle, v_2) E \\ -\partial_{v_1} \mu_-(\langle v \rangle, v_2) E \end{pmatrix},$$

$$R^0_\pm = -\int_0^\infty e^{-\tau \mathcal{A}^0} e^{i\omega\tau} \mathcal{K}^0_\pm E d\tau,$$

$$\rho^0 = \int_{\mathbf{R}^2} (R^0_+ - R^0_-) dv, \qquad j^0_1 = \int_{\mathbf{R}^2} \hat{v}_1 (R^0_+ - R^0_-) dv.$$

Our key estimate is the following.

LEMMA 6. *Let $\mu_\pm$ satisfy* (11) *and* (12), *and let β be any solution of* (4) *of period P. Let* $\operatorname{Im} \omega > 0$. *With ρ and j_1 defined by* (26), (27), *we have*

$$|\rho - \rho^0|_1 + |j_1 - j^0_1|_1 \le C \|\beta\|^{1/2} |E|_1$$

for all $E \in L^1(\mathbf{R}_P)$, where $\|\beta\| = |\beta|_{C^2}$.

We shall show that this lemma follows from the next one.

LEMMA 7. *Under the same conditions,*

$$\int_0^P \left| \int_{\mathbf{R}} \int_0^\infty \int_{\mathbf{R}^2} \{\delta(x - X^\pm) \partial_{v_1} \mu_\pm(e', v'_2) - \delta(x - X^0) \partial_{v_1} \mu_\pm(\langle v' \rangle, v'_2)\} \right.$$

$$\left. \times e^{-\operatorname{Im}\omega\tau} E(x') dv' d\tau dx' \right| dx \le C \|\beta\|^{1/2} |E|_1 \tag{43}$$

for all $E \in L^1(\mathbf{R}_P)$, where $\|\beta\| = |\beta|_{C^1}$, $X^\pm = X^\pm(\tau; 0, x', v')$, and $X^0 = X^0(\tau; 0, x', v')$.

Remark. It is easy to estimate the integral in Lemma 7 by $C|E|_1$, but we will require the small constant $\|\beta\|$. We first illustrate our technique by estimating the free part of the integral as

$$\int_{\mathbf{R}} \int_0^P \int_0^\infty \int_{\mathbf{R}^2} |\delta(x - X^0) \partial_{v'_1} \mu_\pm(\langle v' \rangle, v'_2)| e^{-\operatorname{Im}\omega\tau} |E(x')| dv' d\tau dx dx' \tag{44}$$

$$= \int_0^\infty \int_0^P \int_{\mathbf{R}} \int_{\mathbf{R}^2} e^{-\operatorname{Im}\omega\tau} |E(x')| |\delta(x - x' - \hat{v}'_1 \tau)| |\partial_{v'_1} \mu_+(\langle v' \rangle, v'_2) \hat{v}'_1| dv' dx' dx d\tau$$

$$= \int_0^\infty \int_0^P \int_{x-\tau}^{x+\tau} \int_{\mathbf{R}} e^{-\operatorname{Im}\omega\tau} |E(x') \partial_e \mu_+(\langle \mathcal{V}_0 \rangle, v_2) \hat{\mathcal{V}}_{10}| \langle \mathcal{V}_0 \rangle^3 \tau^{-1} (1 + v_2^2)^{-1} dv_2 dx' dx d\tau,$$

where we have integrated v'_1 first, and computed $\partial_{v'_1}[x - x' - \hat{v}'_1 \tau] = -\tau \langle v' \rangle^{-3} (1 + v_2^2)$. We notice that

$$\int_{x'-\tau}^{x'+\tau} \langle \mathcal{V}_0 \rangle^{3-\gamma} \frac{dx}{\tau} = (1 + v_2^2)^{\frac{3-\gamma}{2}} \int_{x'-\tau}^{x'+\tau} \left(1 - \left| \frac{x - x'}{\tau} \right|^2 \right)^{\frac{\gamma-3}{2}} \tau^{-1} dx$$

$$= (1 + v_2^2)^{\frac{3-\gamma}{2}} \int_{-1}^1 (1 - y^2)^{\frac{\gamma-3}{2}} dy < \infty, \tag{45}$$

since $\gamma > 1$. Thus the free part is bounded by

$$\int_0^\infty e^{-\operatorname{Im}\omega\tau} \int_0^P \int_{x-\tau}^{x+\tau} \int_{\mathbf{R}} |E(x')| \langle \mathcal{V}_0 \rangle^{3-\gamma} \langle v_2 \rangle^{-2-\tilde{\gamma}} \tau^{-1} dv_2 dx' dx d\tau.$$

Since $\tilde\gamma+\gamma>2$ and $\gamma>1$, $\langle v_2\rangle^{1-\tilde\gamma-\gamma}$ is integrable so that we have the bound

$$C\int_0^\infty e^{-\mathrm{Im}\omega\tau}\int_{-\tau}^{P+\tau}|E(x')|dx'd\tau\le C\int_0^\infty e^{-\mathrm{Im}\omega\tau}d\tau(\tau+1)|E|_1\le C|E|_1, \tag{46}$$

where we have used the fact that $\int_{-\tau}^{P+\tau}|E(x')|dx'\le C(\tau+1)|E|_1$.

Proof of Lemma 6. From (31), ρ takes the form

$$\int_{\mathbf{R}}\int_0^\infty\int_{\mathbf{R}^2}\left\{-\delta(x-X^+)\partial_{v_1'}\mu_+(e',v_2')+\delta(x-X^-)\partial_{v_1'}\mu_-(e',v_2')\right\}e^{is\omega}E(x')dv'dsdx'$$

and $\rho^0(x)$ is given by a similar expression. Clearly from Lemma 7, $|\rho-\rho^0|_1\le C\|\beta\|^{1/2}|E|_1$. Similarly $j_1(x)$ takes the form

$$\int_{\mathbf{R}}\int_0^\infty\int_{\mathbf{R}^2}\left\{-\hat V_1^+\delta(x-X^+)\partial_{v_1'}\mu_+(e',v_2')+\hat V_1^-\delta(x-X^-)\partial_{v_1'}\mu_-(e',v_2')\right\}e^{is\omega}E(x')dv'dsdx'$$

and $j_1^0(x)$ is given by a similar expression. Hence (with $j_1=j_{1+}-j_{1-}$, $j_1^0=j_{1+}^0-j_{1-}^0$),

$$\begin{aligned}&\int_0^P|j_{1+}-j_{1+}^0|dx\\&\le\int_0^P\Big|\int_{\mathbf{R}}\int_0^\infty\int_{\mathbf{R}^2}\{\hat V_1^+\left[\delta(x-X^+)\partial_{v_1'}\mu_+(e',v_2')-\delta(x-X^0)\partial_{v_1'}\mu_+(\langle v'\rangle,v_2')\right]\\&\qquad+(\hat V_1^+-\hat V_1^0)\delta(x-X^0)\partial_{v_1'}\mu_+(\langle v'\rangle,v_2')\}e^{-\mathrm{Im}\omega\tau}E(x')dv'd\tau dx'\Big|dx.\end{aligned}$$

Since $|\hat V_1^+|\le1$, the first term is bounded by $C\|\beta\|^{1/2}|E|_1$ because of Lemma 7. On the other hand, since $dV_1^+/d\tau=\beta'(X^+)$, $dV_2^+/d\tau=0$, and $V^0(\tau;0,x',v')=v'$,

$$|\hat V_1^+(s;0,x',v')-\hat v_1'|=\left|\int_0^\tau\beta'(X^+(s;0,x',v'))ds\right|\le\tau\|\beta\|.$$

Thus the second term is bounded by

$$C\|\beta\|\int_0^P\int_{\mathbf{R}}\int_0^\infty\int_{\mathbf{R}^2}\delta(x-X^0)|\partial_{v_1'}\mu_+(\langle v'\rangle,v_2')|\tau e^{-\mathrm{Im}\omega\tau}|E(x')|dv'd\tau dx'dx.$$

This expression is identical to (44) except for the extra factor τ and is therefore estimated in exactly the same way. This proves Lemma 6, assuming Lemma 7.

Proof of Lemma 7. For notational simplicity, we prove only the case of $+$, as the case of $-$ is the same. We split the integral into two major parts according to the time variable. Let

$$T=\min\left\{\frac12|x-x'|^{1/2},\frac1M|x-x'|\right\}\|\beta\|^{-1/2}, \tag{47}$$

where $M=\max\{4,4(P)^{1/2}\}$. We then split (43) as

$$\int_0^P\int_{\mathbf{R}^3}\int_0^\infty d\tau dv'dx'dx=\int_0^P\int_{\mathbf{R}^3}\int_T^\infty+\int_0^P\int_{\mathbf{R}^3}\int_0^T\equiv L+S.$$

The large-time estimate *L*. We further split the integral L into L_1+L_2, where

$$\begin{aligned}L_1&=\int_{\mathbf{R}^3}\int_{\{|x-x'|\ge M^2/4\}\cap[0,P]}\int_T^\infty d\tau dxdx'dv',\\L_2&=\int_{\mathbf{R}^3}\int_{\{|x-x'|<M^2/4\}\cap[0,P]}\int_T^\infty d\tau dxdx'dv'.\end{aligned}$$

We first consider L_1. Since $|x| \le P$ and $|x - x'| \ge 2P$, we are in the untrapped region and it follows that $|x - x'| \ge P/2 + |x'|/2$. Hence

$$T \ge \frac{1}{2}|x - x'|^{1/2}\|\beta\|^{-1/2} \ge c(|x'| + 1)^{1/2}\|\beta\|^{-1/2}.$$

We now employ the exponential decay of $e^{-\mathrm{Im}\omega\tau}$ as

$$\begin{aligned}
L_1 &\le \int_{\mathbf{R}^3} \int_{c(|x'|+1)^{1/2}\|\beta\|^{-1/2}}^{\infty} e^{-\mathrm{Im}\omega\tau}|E(x')| \\
&\times \left\{ \int_0^P \left[\delta(x - X^+)|\partial_{v_1'}\mu_+(e', v_2')| + \delta(x - X^0)|\partial_{v_1'}\mu_+(\langle v'\rangle, v_2')|\right] dx \right\} d\tau dv' dx' \\
&\le C \int_{\mathbf{R}} \int_{c[|x'|+1]^{1/2}\|\beta\|^{-1/2}}^{\infty} e^{-\mathrm{Im}\omega\tau}|E(x')| \int_{\mathbf{R}^2} \left[|\partial_{v_1'}\mu_+(e', v_2')| + |\partial_{v_1'}\mu_+(\langle v'\rangle, v_2')|\right] dv' d\tau dx' \\
&\le C \int_{\mathbf{R}} e^{-c[|x'|+1]^{1/2}\|\beta\|^{-1/2}}|E(x')|dx' \le C \int_{\mathbf{R}} [|x'| + 1]^{-2}\|\beta\|^2|E(x')|dx' \\
&= C\|\beta\|^2 \sum_n \int_{2nP}^{2(n+1)P} [1 + |x'|]^{-2}|E(x')|dx' \le C\|\beta\|^2|E|_1,
\end{aligned} \tag{48}$$

since $\int[|\partial_{v_1'}\mu_+(e', v_2')| + |\partial_{v_1'}(\langle v'\rangle, v_2')|]dv$ is finite because of the decay rate in (11).

For L_2, we have $|x - x'| \le M^2/4$ bounded. Given x and v_2', pick any v_1' such that $X^+(\tau; 0, x', v') = x$. We thus have

$$x - x' = \int_0^\tau \hat{V}_1^+(\theta, x', v')d\theta.$$

By the mean value theorem,

$$\hat{\mathcal{V}}_{10} = \frac{x - x'}{\tau} = \frac{1}{\tau}\int_0^\tau \hat{V}_1^+(\theta; 0, x', v')d\theta = \hat{V}_1^+(s; 0, x', v')$$

for some $s \in [0, \tau]$. Thus $\mathcal{V}_{10} = V_1^+$ and $\mathcal{V}_{20} = v_2' = V_2^+$, so that

$$\langle \mathcal{V}_0\rangle = \langle V^+(s; 0, x', v')\rangle.$$

But by energy conservation along the trajectory, we have

$$\langle v'\rangle + \beta(x') = \langle V^+(s; 0, x', v')\rangle + \beta(X^+(s; 0, x', v')).$$

Therefore

$$|\langle v'\rangle - \langle \mathcal{V}_0\rangle| \le 2\|\beta\|. \tag{49}$$

Since $\tau \ge T = \frac{1}{4}|x - x'|\|\beta\|^{-1/2}$ in the integral L_2, we have

$$|\hat{\mathcal{V}}_{10}| = \left|\frac{x - x'}{\tau}\right| \le 4\|\beta\|^{1/2}$$

so that

$$\langle v'\rangle \le 2\|\beta\| + \langle \mathcal{V}_0\rangle = 2\|\beta\| + \sqrt{\frac{1 + {v_2'}^2}{1 - \hat{\mathcal{V}}_0^2}} \le (1 + C\|\beta\|)\sqrt{1 + {v_2'}^2}.$$

We deduce that $\sqrt{1+{v_1'}^2+{v_2'}^2}-\sqrt{1+{v_2'}^2}\le C\|\beta\|\sqrt{1+{v_2'}^2}$, so that

$$ {v_1'}^2 \le C\|\beta\|\left[\sqrt{1+{v_1'}^2+{v_2'}^2}\sqrt{1+{v_2'}^2}+1+{v_2'}^2\right]\le \frac{C}{2}\|\beta\|\left[{v_1'}^2+5({v_2'}^2+1)\right] $$

and

$$ |v_1'|\le C\|\beta\|^{1/2}\sqrt{1+{v_2'}^2}\equiv m. $$

Therefore, employing the smallness of v_1' and the boundedness of x', we get the term L_2 bounded by

$$ \int_0^\infty\int_{-C}^{C}\int_{\mathbf{R}}\int_{-m}^{m}\left\{\int_0^P|\partial_e\mu_+||\hat v_1'|\delta(x-X^+)dx\right\}|E(x')|e^{-\mathrm{Im}\omega\tau}dv_1'dv_2'dx'd\tau $$

plus the same term with X^+ replaced by X^0. Therefore, we get the bound

$$ \begin{aligned} &C|E|_1\int_0^\infty\int_{\mathbf{R}}\int_{-m}^{m}\langle v_2'\rangle^{-\tilde\gamma-\gamma}dv_1'dv_2'e^{-\mathrm{Im}\omega\tau}d\tau\\ &\le C|E|_1\|\beta\|^{1/2}\int_0^\infty e^{-\mathrm{Im}\omega\tau}d\tau\le C\|\beta\|^{1/2}|E|_1, \end{aligned} $$

since $\tilde\gamma+\gamma>1$. Combining this with (48), we have the large-time estimate

$$ L=L_1+L_2\le C|E|_1\|\beta\|^{1/2}. \tag{50} $$

The small-time estimate *S*. Now in (43) we consider the case $\tau\le T$. Since

$$ 0\le\tau\le\frac{1}{2}|x-x'|^{1/2}\|\beta\|^{-1/2}, $$

we have

$$ \left|\frac{x-x'}{\tau}\right|\ge 2|x-x'|^{1/2}\|\beta\|^{1/2}\ge 4\tau\|\beta\|\ge 4\left|\frac{1}{\tau}\int_0^\tau\int_0^\theta\beta'(X^+(s))dsd\theta\right|. $$

Now for any v' such that $x=X^+(\tau;0,x',v')$, we have the trajectory equation

$$ x=x'+\hat v_1'\tau\pm\int_0^\tau\int_0^\theta\langle V^+\rangle^{-3}\beta'(X^+)dsd\theta. $$

Thus,

$$ \begin{aligned} &|\hat v_1'|\ge\left|\frac{x-x'}{\tau}\right|-\left|\frac{1}{\tau}\int_0^\tau\int_0^\theta\beta'(X^+(s))dsd\theta\right|\ge\frac{3}{4}\left|\frac{x-x'}{\tau}\right|\ge 3\|\beta\|^{1/2},\\ &|\hat v_1'|\le\left|\frac{x-x'}{\tau}\right|+\left|\frac{1}{\tau}\int_0^\tau\int_0^\theta\beta'(X^+(s))dsd\theta\right|\le\frac{5}{4}\left|\frac{x-x'}{\tau}\right|. \end{aligned} \tag{51} $$

We have used the fact $\tau\le\frac{1}{4}|x-x'|\|\beta\|^{-1/2}$ by the definition of T in (47). Therefore, v' must lie in the *untrapped region*. Indeed, by (37), in the trapped region all the

velocities are bounded by $2\|\beta\|^{1/2}$ and this contradicts (51). By Lemma 5, only one such $\mathcal{V}_{1+}$ exists. Hence,

$$\int_{\mathbf{R}} \partial_{v_1'}\mu_+(e', v_2')\delta(x - X^+)dv_1' = \hat{\mathcal{V}}_{1+}\partial_e\mu_+'(W_+, v_2')\mathcal{J},$$

where $W_+ = \langle\mathcal{V}_+\rangle - \beta(X^+)$ and

$$\mathcal{J} = \left\{ \left[\frac{\partial X^+}{\partial v_1'}\right]\Big|_{v_1'=\mathcal{V}_{1+}} \right\}^{-1}.$$

In the free case, $\hat{\mathcal{V}}_{10} = (x - x')/\tau$, and $(\partial X_0/\partial v_1')^{-1} = \tau^{-1}\langle\mathcal{V}_0\rangle^3$. Therefore, by first integrating over the v' variable, we get the small-time integral equal to

$$\begin{aligned} S &= \int |E(x')|e^{-\mathrm{Im}\omega\tau}|\hat{\mathcal{V}}_{1+}\partial_e\mu_+(W_+, v_2')\mathcal{J} - \hat{\mathcal{V}}_{10}\partial_e\mu_+(\langle\mathcal{V}_0\rangle, v_2')\tau^{-1}\langle\mathcal{V}_0\rangle^3| \\ &\le \int |E(x')|e^{-\mathrm{Im}\omega\tau}|\partial_e\mu_+(W_+, v_2')\{\mathcal{J} - (1 + {v_2'}^2)^{-1}\hat{\mathcal{V}}_{1+}(x - x')^{-1}\langle\mathcal{V}_+\rangle^3\}| \\ &+ \int \frac{|E(x')|}{|x - x'|(1 + {v_2'}^2)}e^{-\mathrm{Im}\omega\tau}|\partial_e\mu_+(W_+, v_2')\hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3 - \partial_e\mu_+(\langle\mathcal{V}_0\rangle, v_2')\hat{\mathcal{V}}_{10}^2\langle\mathcal{V}_0\rangle^3| \\ &\equiv S_1 + S_2, \end{aligned} \tag{52}$$

where $\int = \int_0^P \int_{\mathbf{R}} \int_0^T \int_{\mathbf{R}} dv_2' d\tau dx' dx$.

We now estimate S_1. In order to compute $\mathcal{J}$, we take the v_1' derivative in (38) to get (for $t > 0$)

$$\begin{aligned} 0 =& \frac{\partial X^+}{\partial v_1'}\frac{\langle v'\rangle - \beta(x') + \beta(X^+(v'))}{\{[\langle v'\rangle - \beta(x') + \beta(X^+(v'))]^2 - 1 - v_2^2\}^{1/2}} \\ &- \int_{x'}^{X^+(v')} [(\langle v'\rangle - \beta(x') + \beta(y))^2 - 1 - v_2^2]^{-3/2}\hat{v}_1'(1 + v_2^2)dy. \end{aligned}$$

Put $v' = \mathcal{V}_+(\tau, x, x')$, so that $X^+(\tau, 0, x', v') = x$ and

$$\mathcal{J}^{-1} = (1 - (1 + v_2^2)A(x)^{-2})^{1/2}\hat{\mathcal{V}}_{1+}Q(1 + v_2^2),$$

where $A(y) = \langle\mathcal{V}_+(\tau, x, x')\rangle - \beta(x') + \beta(y)$, $\mathcal{V}_+ = \mathcal{V}_+(\tau, x, x')$, and

$$Q = \int_{x'}^{x} (A(y)^2 - 1 - v_2^2)^{-3/2}dy.$$

From (51), we have

$$|\hat{\mathcal{V}}_{1+}| \ge \frac{3}{4}\left|\frac{x - x'}{\tau}\right| \ge 3\|\beta\|^{1/2}. \tag{53}$$

To estimate Q, notice that $|A(y) - \langle\mathcal{V}_+\rangle| \le 2\|\beta\|$ and therefore $A(y)$ is comparable to $\langle\mathcal{V}_+\rangle$. Hence,

$$\begin{aligned} A^2 - 1 - v_2^2 &= \mathcal{V}_{1+}^2 + (A - \langle\mathcal{V}_+\rangle)(A + \langle\mathcal{V}_+\rangle) \\ &\ge \mathcal{V}_{1+}^2 - 2\|\beta\|(2\langle\mathcal{V}_+\rangle + 2\|\beta\|) \\ &\ge \mathcal{V}_{1+}^2 - \frac{4}{3}\|\beta\|^{1/2}|\mathcal{V}_{1+}| - 4\|\beta\|^2 \\ &\ge \frac{5}{9}\mathcal{V}_{1+}^2 - 4\|\beta\|^2 \ge \frac{1}{2}\mathcal{V}_{1+}^2 \end{aligned}$$

by (53) and the smallness of β. Letting $g(A) = (A^2 - 1 - v_2^2)^{-3/2}$, we have

$$|g'(A)| = 3(A^2 - {v_2'}^2 - 1)^{-5/2} A \le C|\mathcal{V}_{1+}|^{-5}\langle \mathcal{V}_+\rangle.$$

Therefore, we have

$$|Q(\tau, x, x') - (x - x')|\mathcal{V}_{1+}|^{-3}| \le C\|\beta\||\mathcal{V}_{1+}|^{-5}\langle \mathcal{V}_+\rangle |x - x'|$$

so that Q is comparable to $(x - x')|\mathcal{V}_{1+}|^{-3}$ and

$$\begin{aligned}
&|Q^{-1}(\tau, x, x') - (x - x')^{-1}|\mathcal{V}_{1+}|^3| \\
&\le C\|\beta\||\mathcal{V}_{1+}|^{-5}\langle \mathcal{V}_+\rangle |x - x'|\{(x - x')^{-1}|\mathcal{V}_{1+}|^3\}^2 \\
&= C\|\beta\||x - x'|^{-1}|\mathcal{V}_{1+}|\langle \mathcal{V}_+\rangle.
\end{aligned}$$

Also, inside $\mathcal{J}$ we have

$$(1 - (1 + v_2^2)A^{-2})^{-1/2} = \langle \mathcal{V}_+\rangle |\mathcal{V}_{1+}|^{-1} + O[\|\beta\|(1 + v_2^2)|\mathcal{V}_{1+}|^{-3}].$$

We thus estimate $\mathcal{J}$ as

$$\begin{aligned}
&|\mathcal{J} - (1 + v_2^2)^{-1}\langle \mathcal{V}_+\rangle^2 \mathcal{V}_{1+}(x - x')^{-1}| \\
&= \frac{1}{(1 + v_2^2)}\left|(1 - (1 + v_2^2)^{-1}A^{-2})^{-1/2} \cdot \hat{\mathcal{V}}_{1+}^{-1} Q^{-1} - \langle \mathcal{V}_+\rangle |\mathcal{V}_{1+}|^{-1} \cdot \hat{\mathcal{V}}_{1+}^{-1}|\mathcal{V}_{1+}|^3 (x - x')^{-1}\right| \\
&\le C\|\beta\||x - x'|^{-1}|\hat{\mathcal{V}}_{1+}^{-1}|[1 + \langle \mathcal{V}_+\rangle^2 (1 + v_2^2)^{-1}] \\
&\le C\|\beta\|^{1/2}\tau^{-1}\hat{\mathcal{V}}_{1+}^{-1}\langle \mathcal{V}_+\rangle^2 (1 + v_2^2)^{-1}.
\end{aligned}$$

In the last step, we have used the fact that $\tau|x - x'|^{-1} \le \frac{1}{4}\|\beta\|^{-1/2}$ from (47). Moreover, from the trajectory equation and (51), (49) with $v' = \mathcal{V}_+$, we have

$$\frac{3}{4}\left|\frac{x - x'}{\tau}\right| \le |\hat{\mathcal{V}}_{1+}| \le \frac{5}{4}\left|\frac{x - x'}{\tau}\right|, \quad |\hat{\mathcal{V}}_{1+} - \hat{\mathcal{V}}_{10}| \le \tau\|\beta\|.$$

Using the energy conservation along the trajectory, we estimate $W_+ \equiv \langle \mathcal{V}_+\rangle - \beta(x')$ as $|W_+ - \langle \mathcal{V}_+\rangle| \le \|\beta\|$. Since $\langle \mathcal{V}_0\rangle$ dominates $\|\beta\|$, we also have

$$c\langle \mathcal{V}_0\rangle \le \langle \mathcal{V}_+\rangle \le C\langle \mathcal{V}_0\rangle. \tag{54}$$

We therefore estimate S_1 by

$$\begin{aligned}
&\int_0^P \int_{x-\tau}^{x+\tau} \int_0^T \int_{\mathbf{R}} |E(x')| e^{-\mathrm{Im}\omega\tau} \left|\partial_e \mu_+(W_+, v_2')\hat{\mathcal{V}}_{1+}\left\{\mathcal{J} - (1 + v_2^2)^{-1}\hat{\mathcal{V}}_{1+}(x - x')^{-1}\langle \mathcal{V}_+\rangle^3\right\}\right| \\
&\le C\int_0^P \int_{x-\tau}^{x+\tau} \int_0^T \int_{\mathbf{R}} e^{-\mathrm{Im}\omega\tau}|\partial_e \mu_+(W_+, v_2')\left[\|\beta\|^{1/2}\tau^{-1}\langle \mathcal{V}_+\rangle^2\right]|E(x')|(1 + {v_2'}^2) \qquad (55)
\end{aligned}$$

with the volume element $dv_2' d\tau dx' dx$. In this integral we have

$$|\partial_e \mu_+(W_+, v_2')|\langle \mathcal{V}_+\rangle^2 \le C\langle \mathcal{V}_+\rangle^{2-\gamma} \le C\langle \mathcal{V}_0\rangle^{2-\gamma}$$

with $\gamma > 1$ so that

$$S_1 \le C\|\beta\|^{1/2}|E|_1$$

just as in (46).

From (54) and the decay condition in (11), we now estimate the main part of the integrand in S_2 as

$$\begin{aligned} J_2 \equiv & |\partial_e\mu_+(W_+, v_2')\hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3 - \partial_e\mu_+(\langle\mathcal{V}_0\rangle, v_2')\hat{\mathcal{V}}_{10}^2\langle\mathcal{V}_0\rangle^3| \\ & \le \left|[\partial_e\mu_+(W_+, v_2') - \partial_e\mu_+(\langle\mathcal{V}_0\rangle, v_2')]\,\hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3\right| \\ & + \left|\partial_e\mu(W_0, v_2')[\hat{\mathcal{V}}_{10}^2\langle\mathcal{V}_0\rangle^3 - \hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3]\right|. \end{aligned}$$

Notice that

$$\begin{aligned} & |\hat{\mathcal{V}}_{10}^2\langle\mathcal{V}_0\rangle^3 - \hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3| = |\mathcal{V}_{10}^2\langle\mathcal{V}_0\rangle - \mathcal{V}_{1+}^2\langle\mathcal{V}_+\rangle| \\ & = |\langle\mathcal{V}_0\rangle - \langle\mathcal{V}_+\rangle|(\langle\mathcal{V}_0\rangle^2 + \langle\mathcal{V}_0\rangle\langle\mathcal{V}_+\rangle + \langle\mathcal{V}_+\rangle^2 - 1 - {v_2'}^2) \\ & \le C|\langle\mathcal{V}_0\rangle - \langle\mathcal{V}_+\rangle|\langle\mathcal{V}_0\rangle^2. \end{aligned}$$

By (49) and (11), we thus estimate the main part of the integrand in S_2 by

$$\begin{aligned} J_2 \le & \{\sup_{\theta\in[\langle\mathcal{V}_0\rangle, W_+]} \partial_{ee}\mu_+(\theta, v_2')|[W_+ - \langle\mathcal{V}_0\rangle]\hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3| \\ & + C|\partial_e\mu_+(\langle\mathcal{V}_0\rangle, v_2')||\langle\mathcal{V}_0\rangle - \langle\mathcal{V}_+\rangle|\langle\mathcal{V}_0\rangle^2\} \\ & \le C\|\beta\|\langle\mathcal{V}_0\rangle^{3-\gamma}\langle v_2'\rangle^{-\tilde{\gamma}}. \end{aligned}$$

From the definition of T in (47), we have $|x - x'|^{-1} \le \frac{1}{4\tau}\|\beta\|^{-1/2}$. We plug this into S_2 in (52) to get

$$\begin{aligned} S_2 = & \int_0^P \int_0^T \int_{x-\tau}^{x+\tau} \int_{\mathbf{R}} e^{-\mathrm{Im}\omega\tau} \frac{|E(x')|}{|x - x'|(1 + {v_2'}^2)} \\ & \times \left|\partial_e\mu_+(W_+, v_2')\hat{\mathcal{V}}_{1+}^2\langle\mathcal{V}_+\rangle^3 - \partial_e\mu_+(\langle\mathcal{V}_0\rangle, v_2')\hat{\mathcal{V}}_{10}^2\langle\mathcal{V}_0\rangle^3\right| dv_2'dx'd\tau dx \\ & \le C\|\beta\|^{1/2}\int_0^P \int_0^T \int_{x-\tau}^{x+\tau} \int_{\mathbf{R}} e^{-\mathrm{Im}\omega\tau}|E(x')|\langle\mathcal{V}_0\rangle^{3-\gamma}\langle v_2'\rangle^{-2-\tilde{\gamma}}\tau^{-1}dv_2'dx'd\tau dx \\ (56) \ & \le C\|\beta\|^{1/2}|E|_1, \end{aligned}$$

where we have again used (45) to integrate over x. Thus we deduce the small-time estimate

$$S \le S_1 + S_2 \le C\|\beta\|^{1/2}|E|_1. \tag{57}$$

The lemma follows from (50) and (57). □

LEMMA 8. *The same estimate as in Lemma* 7 *is valid if* $\exp[-\mathrm{Im}\omega\tau]$ *is replaced by* $\tau^m \exp[-\mathrm{Im}\omega\tau]$ *for any* $m \ge 1$.

The proof is identical to the preceding one.

Now we are ready for our main theorem about the linear operator $\mathcal{C}$. Recall the definition of $\mathcal{C}(\omega, \beta)$ in (25). We shall write $P = 2P_\beta$. Furthermore, we define $\mathcal{C}(\omega, 0)$ from $L^1(\mathbf{R}_P)$ into itself by

$$\mathcal{C}(\omega, 0)E(x) = \int_0^x \rho^0(y)dy + \frac{1}{i\omega P}\int_0^P j_1^0(y)dy - \frac{1}{P}\int_0^P \int_0^z \rho^0(y)dydz. \tag{58}$$

We define the closely related operator $\mathcal{C}_0$ from $L^1(\mathbf{R}_{2P_0})$ to itself, to be given by the same formula but acting on functions in $L^1(\mathbf{R}_{2P_0})$ and with P replaced by $2P_0$.

THEOREM 2 (growing mode for periodic BGK equilibria). *Let $P = 2P_\beta$ and $\operatorname{Im}\omega > 0$ and $\gamma > 1$, and $\mu_\pm$ satisfy* (11), (12), *and* (14). *Then*

(a) *$\mathcal{C}(\omega,\beta)$ and $\mathcal{C}(\omega,0)$ are compact operators from $L^1(\mathbf{R}_P)$ to $L^1(\mathbf{R}_P)$ such that*

$$\|\mathcal{C}(\omega,\beta) - \mathcal{C}(\omega,0)\|_{L^1(\mathbf{R}_P)\to L^1(\mathbf{R}_P)} \le C\|\beta\|_{C^1}^{1/2},$$

where $\mathcal{C}(\omega,0)$ is the unperturbed linearized operator. The constant C is uniform for $\operatorname{Im}\omega >$ constant > 0.

(b) *$\mathcal{C}(\omega,\beta)$ is analytic in ω for $\operatorname{Im}\omega > 0$.*

(c) *There exists $\eta > 0$ such that if $\|\beta\|_{C^1} < \eta$, there exists a growing mode $[g^\pm, E]$ with period P for the linearized Vlasov–Maxwell system* (19) *around $[\mu_\pm(\langle v\rangle \mp \beta(x), v_2), \beta']$.*

Proof. By definition

$$(\mathcal{C}(\omega,\beta)E - \mathcal{C}(\omega,0)E)(x) = \int_0^x [\rho(y) - \rho^0(y)]dy + \frac{1}{i\omega P}\int_0^P [j_1(y) - j_1^0(y)]dy - \frac{1}{P}\int_0^P\int_0^z [\rho(y) - \rho^0(y)]dydz.$$

By Lemma 6,

$$|\mathcal{C}(\omega,\beta)E - \mathcal{C}(\omega,0)E|_1 \le C\|\beta\|^{1/2}|E|_1$$

where C may depend on P and ω. Now by (32)

$$|\partial_x\{\mathcal{C}(\omega,\beta)E\}|_1 = |\rho|_1 \le C|E|_1$$

and, by definition of $\mathcal{C}(\omega,\beta)$,

$$|\mathcal{C}(\omega,\beta)E|_1 \le C|\rho|_1 + C|j_1|_1/|\omega| \le C|E|_1.$$

Since $W^{1,1}(\mathbf{R}_P)$ is compact in $L^1(\mathbf{R}_P)$, $\mathcal{C}(\omega,\beta)$ is a compact operator. This proves (a).

For part (b), notice that $\mathcal{C}(\omega,\beta)E$ is given by an absolutely convergent integral, in which ω appears as $e^{i\omega\tau}$. The integral converges uniformly in each half-plane $\{\operatorname{Im}\omega \ge c > 0\}$. By Lemma 8 with $m = 1$ and by repeating the argument in part (a), we get

$$\left\|\frac{d}{d\omega}[\mathcal{C}(\omega,\beta) - \mathcal{C}(\omega,0)]\right\| \le C\|\beta\|^{1/2}.$$

We thus deduce part (b).

To prove part (c), we define the dilation operator G_β from $L^1(\mathbf{R}_{2P_0})$ to $L^1(\mathbf{R}_P)$ as $G_\beta : E(x) \to E(x/\lambda)$ where $\lambda = P_\beta/P_0$. Clearly G_β is a one-to-one and bounded linear operator from $L^1(\mathbf{R}_{2P_0})$ to $L^1(\mathbf{R}_P)$. We *claim that*

$$\|G_\beta^{-1}\mathcal{C}(\omega,\beta)G_\beta - \mathcal{C}_0\|_{L^1\to L^1} \le C\|\beta\|^{1/2}$$

for $\|\beta\|$ small.

Proof of the claim. Notice that

$$\begin{aligned}
&\|G_\beta^{-1}\mathcal{C}(\omega,\beta)G_\beta - \mathcal{C}_0\| \\
&\le \|G_\beta^{-1}\{\mathcal{C}(\omega,\beta)-\mathcal{C}(\omega,0)\}G_\beta\| + \|G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta - \mathcal{C}_0\| \\
&\le \|G_\beta^{-1}\|\|\mathcal{C}(\omega,\beta)-\mathcal{C}(\omega,0)\|\|G_\beta\| + \|G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta - \mathcal{C}_0\| \\
&\le C\|\beta\|^{1/2} + \|G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta - \mathcal{C}_0\|,
\end{aligned}$$

where we have used part (a). Now we consider $G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta$. Notice that by changing variables $x'' = x'/\lambda$ and $y = \lambda y'$, we have by (28), (31), and (42)

$$\begin{aligned}
G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta E(x) &= \int_0^{\lambda x}\int_{\mathbf{R}} E(x'/\lambda)k_0(y,x')dx'dy + \frac{1}{i\omega P}\int_0^P j_1^0(x'/\lambda)dx' \\
&\quad -\frac{1}{P}\int_0^P\int_0^z\int_{\mathbf{R}} E(x'/\lambda)k_0(y,x')dx'dydz \\
&= \int_0^x\int_{\mathbf{R}} E(x'')k_0(\lambda y',\lambda x'')dx''dy' + \frac{1}{2i\omega P_0}\int_0^{2P_0} j_1^0(x'')dx'' \\
&\quad -\frac{1}{2P_0}\int_0^{2P_0}\int_0^z\int_{\mathbf{R}} E(x'')k_0(\lambda y',\lambda x'')dx''dy'dz.
\end{aligned}$$

Therefore we have

$$|\partial_x(G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta E) - \partial_x(\mathcal{C}_0E)|_{L^1} \le \int_0^{2P_0}\int_{\mathbf{R}} |E(x')||k_0(\lambda x,\lambda x') - k_0(x,x')|dx'dx.$$

We notice that

$$k_0^\pm(x,x',\omega) = \mp\int_0^\infty \hat{k}_0^\pm(x,x',t)e^{i\omega t}dt,$$

where $\hat{k}_0^\pm(x,x',t) = \int_{\mathbf{R}^2}\delta(x-x'-\hat{v}_1't)\partial_{v_1'}\mu_\pm(\langle v'\rangle, v_2')dv'$. Hence,

$$\begin{aligned}
&k_0^\pm(\lambda x,\lambda x',\omega) = \mp\int_0^\infty \hat{k}_0^\pm(\lambda x,\lambda x',t)e^{i\omega t}dt = \mp\int_0^\infty \hat{k}_0^\pm(\lambda x,\lambda x',\lambda s)e^{i\omega\lambda s}\lambda ds \\
&= \mp\int_0^\infty \hat{k}_0^\pm(x,x',s)e^{i\omega\lambda s}ds = k_0^\pm(x,x',\lambda\omega)
\end{aligned}$$

since $\hat{k}_0^\pm(\lambda x,\lambda x',\lambda t) = \lambda^{-1}\hat{k}_0^\pm(x,x',t)$. Therefore,

$$\begin{aligned}
&k_0^\pm(\lambda x,\lambda x',\omega) - k_0^\pm(x,x',\omega) = k_0^\pm(x,x',\lambda\omega) - k_0^\pm(x,x',\omega) \\
&= \mp\int_0^\infty \hat{k}_0^\pm(x,x',t)[e^{i\lambda\omega t} - e^{i\omega t}]dt \\
&\le |\lambda-1|\int_0^\infty |\hat{k}_0^\pm(x,x',t)||\omega t|e^{-\lambda^* t\mathrm{Im}\omega}dt,
\end{aligned}$$

where $\lambda^* = \max(\lambda,1)$. Hence as in (44)–(46),

$$\begin{aligned}
&|\partial_x(G_\beta^{-1}\mathcal{C}(\omega,0)G_\beta E) - \partial_x(\mathcal{C}_0E)|_{L^1} \\
&\le |\lambda-1||\omega|\int_0^{2P_0}\int_{\mathbf{R}}\int_0^\infty |E(x')||\hat{k}_0^\pm(x,x',t)||t|e^{-\lambda^* t\mathrm{Im}\omega}dtdx'dx \\
&\le |\lambda-1|C|E|_{L^1(\mathbf{R}_{2P_0})}.
\end{aligned}$$

The claim is proved.

We define for $0 \le s \le 1$, $T(\omega, s) = \mathcal{C}_0 + s(G_\beta^{-1}\mathcal{C}(\omega,\beta)G_\beta - \mathcal{C}_0)$. By parts (a) and (b), $I - T(\omega, s)$ is a compact operator from $L^1(\mathbf{R}_{2P_0})$ to itself, is analytic in ω, and is continuous in s. By Lemma 2, $I - \mathcal{C}_0$ has a nontrivial nullspace, so that there is a pole ω_0, fixed, of $(I - \mathcal{C}_0)^{-1} = (I - T(\omega, 0))^{-1}$ with $\operatorname{Im}\omega_0 > 0$. We choose $\epsilon_0 > 0$ so small that for $|\omega - \omega_0| = \epsilon_0$, the operator $(I - T(\omega, 0))$ is invertible and $\operatorname{Im}\omega > 0$. If $\|\beta\|_{C^1} < \eta$ is small enough, then from the claim

$$\|T(\omega, s) - T(\omega, 0)\| \le s\|G_\beta^{-1}\mathcal{C}(\omega,\beta)G_\beta - \mathcal{C}_0\| \le C\eta^{1/2}.$$

Hence $I - T(\omega, s)$ is also invertible on $|\omega - \omega_0| = \epsilon_0$ for all $0 \le s \le 1$. Since the poles of $(I - T(\omega, s))^{-1}$ are continuous in s, as is well known [St], there is a pole of $(I - T(\omega, 1))^{-1}$ in $|\omega - \omega_0| < \epsilon_0$. In particular, let ω_β be a pole of $(I - T(\omega, 1))^{-1} = (I - G_\beta^{-1}\mathcal{C}(\omega,\beta)G_\beta)$ in $|\omega - \omega_0| < \epsilon_0$. Then $\operatorname{Im}\omega_\beta > 0$ and $G_\beta^{-1}\mathcal{C}(\omega_\beta,\beta)G_\beta$ has the eigenvalue 1. Hence there exists an $E_0 \ne 0$ and $E_0 \in L^1(\mathbf{R}_{2P_0})$ such that $G_\beta^{-1}\mathcal{C}(\omega_\beta,\beta)G_\beta E_0 = E_0$. Therefore

$$\mathcal{C}(\omega_\beta,\beta)G_\beta E_0 = G_\beta E_0.$$

By Lemma 4, part (b), we complete the proof. □

4. Properties of periodic eigenfunctions. The following lemma of Vidav [V, Sh] will be used to obtain the linearized estimate.

LEMMA 9. *Let Y be a Banach space and A be a linear operator that generates a strongly continuous semigroup on Y such that $\|e^{-tA}\| \le Me^{\alpha t}$ for all $t \ge 0$. Let K be a compact operator from Y to Y. Then $A + K$ generates a strongly continuous semigroup $e^{-t(A+K)}$, and the spectrum of $(-A - K)$ consists of a finite number of eigenvalues of finite multiplicity in $\{\operatorname{Re}\lambda > \delta\}$ for every $\delta > \alpha$. These eigenvalues can be labeled by*

$$\operatorname{Re}\lambda_1 \ge \operatorname{Re}\lambda_2 \ge \cdots \ge \operatorname{Re}\lambda_n \ge \delta.$$

Furthermore, for every $\Lambda > \operatorname{Re}\lambda_1$, there is a constant C_Λ such that

$$\|e^{-t(A+K)}\|_{L(Y,Y)} \le C_\Lambda e^{\Lambda t}.$$

Applying this lemma and Lemma 3 to the linearized periodic Vlasov–Maxwell system $1\frac{1}{4}$L in (8), we deduce the following.

LEMMA 10 (linear Vlasov–Maxwell). *Let $\mu_\pm$ satisfy* (11) *and* (12), *and let β be any solution of* (4) *of period P. Then for all $\delta > 0$, the spectrum of $-\mathcal{L}$ in $\{\operatorname{Re}\lambda > \delta\}$ consists of a finite number of eigenvalues of finite multiplicity. If λ_1 denotes the eigenvalue of the maximal real part, and $\Lambda > \max\{0, \operatorname{Re}\lambda_1\}$, then there exists $C_\Lambda > 0$ such that*

$$\|e^{-t\mathcal{L}}u_0\|_m \le C_\Lambda e^{\Lambda t}\|u_0\|_m.$$

Proof. We apply the previous lemma to the space $Y = \mathcal{M}$ and the operator $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$ given by (23). By (24) we may take any $\delta > 0$. □

LEMMA 11 (regularity of eigenfunctions). *Let $\mu_\pm$ satisfy* (11) *and* (12), *and let β be any solution of* (4) *of period P. Let λ be any eigenvalue of $-\mathcal{L}$ with $\operatorname{Re}\lambda > 0$ and $[R_\pm, E_0]$ its eigenfunction triple. Assume $\|\beta\| < (\operatorname{Re}\lambda)^2$. Then $R_\pm \in W^{1,1}(\mathbf{R}_P \times \mathbf{R}^2)$ and there exists a constant C depending only on λ and $\mu_\pm$ such that*

$$|E_0|_{1,1} + \|R\|_{1,1} \le C\|R\|_1.$$

Proof. We begin with $u = [R_\pm, E_0] \in \mathcal{M}$. We first claim that

$$R = -\int_0^\infty e^{-t\mathcal{A}} e^{-\lambda t} \mathcal{K}(E_0) dt. \tag{59}$$

In order to prove this, notice that $g(t) = e^{\lambda t} R$ and $E(t) = e^{\lambda t} E_0$ satisfy $\partial_t g + \mathcal{A} g = -\mathcal{K} E$. Hence

$$e^{\lambda t} R = e^{-(t-s)\mathcal{A}} e^{\lambda s} R + \int_s^t e^{-(t-\tau)\mathcal{A}} \mathcal{K} e^{\lambda \tau} E_0 d\tau.$$

Letting $s \to -\infty$, we get

$$e^{\lambda t} R = -\int_{-\infty}^t e^{-(t-\tau)\mathcal{A}_\pm} \mathcal{K} E_0 e^{\lambda \tau} d\tau = -\int_0^\infty e^{-\tau \mathcal{A}} \mathcal{K} E_0 e^{\lambda(t-\tau)} d\tau,$$

which is the same as (59). The integral converges because $\mathrm{Re}\lambda > 0$.

Since $[R_\pm, E_0] \in \mathcal{M}$, we have $\partial_x E_0 = \rho = \int (R_+ - R_-) dv$ so that $E_0 \in L^1(\mathbf{R}_P)$ and $\mathcal{K}(E_0) \in L^1(\mathbf{R}_P \times \mathbf{R}^2)$. Writing $\exp(-t\mathcal{A})$ in terms of the characteristics as in the proof of Lemma 4(c), we see that $\exp(-t\mathcal{A})$ also maps L^1 into itself and $W^{1,1}$ into itself. So (59) implies that $R \in L^1(\mathbf{R}_P \times \mathbf{R}^2)$. Hence $\rho \in L^1(\mathbf{R}_P)$ and

$$|E_0|_{1,1} \le C \|R\|_1.$$

Next we let $h(t) = \exp(-t\mathcal{A})(\mathcal{K} E_0)$. Since $\mathcal{K} E_0 \in W^{1,1}$, it follows that $h(t) \in W^{1,1}$. Thus $(\partial_t + \mathcal{A}) h = 0, h(0) = \mathcal{K} E_0$. Differentiating this equation with respect to x, we get

$$(\partial_t + \mathcal{A})(\partial_x h) = [\mathcal{A}, \partial_x] h = \begin{pmatrix} -\beta'' & 0 \\ 0 & \beta'' \end{pmatrix} \partial_{v_1} h, \quad \partial_x h(0) = \partial_x \mathcal{K} E_0,$$

where $[\mathcal{A}, \partial_x]$ is the commutator. Hence

$$\partial_x h(t) = e^{-t\mathcal{A}} \partial_x \mathcal{K} E_0 + \int_0^t e^{-(t-\tau)\mathcal{A}} \begin{pmatrix} -\beta'' & 0 \\ 0 & \beta'' \end{pmatrix} \partial_{v_1} h(\tau) d\tau.$$

Similarly,

$$\partial_{v_1} h(t) = e^{-t\mathcal{A}} \partial_{v_1} \mathcal{K} E_0 - \int_0^t e^{-(t-\tau)\mathcal{A}} \langle v \rangle^{-3} (1 + v_2^2) \partial_x h(\tau) d\tau,$$
$$\partial_{v_2} h(t) = e^{-t\mathcal{A}} \partial_{v_2} \mathcal{K} E_0 + \int_0^t e^{-(t-\tau)\mathcal{A}} v_1 v_2 \langle v \rangle^{-3} \partial_x h(\tau) d\tau.$$

Taking L^1-norms, we get

$$\|\partial_x h(t)\|_1 + \|\partial_v h(t)\|_1 \le (\|\partial_x (\mathcal{K} E_0)\|_1 + \|\partial_v (\mathcal{K} E_0)\|_1) e^{t \|\beta\|_{C^2}^{1/2}}.$$

We put this estimate into the integrand of (59) to get

$$\|\partial_x R\|_1 + \|\partial_v R\|_1 \le \left(\int_0^\infty e^{[\|\beta\|_{C^2}^{1/2} - \lambda] t} dt \right) (\|\partial_x \mathcal{K} E_0\|_1 + \|\partial_v \mathcal{K} E_0\|_1).$$

By the definition of (21) of $\mathcal{K}$, the decay condition (11) on $\mu_\pm$, and the boundedness of β, we deduce that $\mathcal{K} E_0 \in W^{1,1}$, and we have the desired estimate. ∎

LEMMA 12. *Let $\underline{u}$ be an eigenvector of $-\mathcal{L}$ with its eigenvalue λ, $\mathrm{Re}\lambda > 0$. If λ is not real, then there is a constant $\zeta > 0$ such that for all $t > 0$*

$$\|e^{-\mathcal{L}t}(\mathrm{Im}\ \underline{u}\ )\|_1 \geq \zeta e^{\mathrm{Re}\lambda t}\|\mathrm{Im}\ \underline{u}\ \|_1 > 0.$$

Proof. We prove it by contradiction. Notice that

$$e^{-\mathcal{L}t}(\mathrm{Im}\ \underline{u}\ ) = \mathrm{Im}(e^{-\mathcal{L}t}\underline{u}) = e^{\mathrm{Re}\lambda t}(\sin[\mathrm{Im}\ \lambda t]\mathrm{Re}\ \underline{u}\ + \cos[\mathrm{Im}\ \lambda t]\mathrm{Im}\ \underline{u}\ ).$$

If the lemma were false, by passing through a convergent subsequence of $\sin[\mathrm{Im}\ \lambda t_n]$, and $\cos[\mathrm{Im}\ \lambda t_n]$ with $n \to \infty$, we would have $a\mathrm{Im}\ \underline{u}\ + b\mathrm{Re}\ \underline{u}\ = 0$, with $a^2 + b^2 = 1$. Therefore either $\mathrm{Im}\ \underline{u}$ or $\mathrm{Re}\ \underline{u}$ would be a real eigenvector and λ would be real, a contradiction. ∎

LEMMA 13 (pointwise estimate of eigenfunctions). *Let $\mu_\pm$ satisfy* (11), (12) *and let β be any solution of* (4) *of period P. Let $[R_+, R_-, E_0]$ be an eigenvector with $\|R\|_{1,1} + |E_0|_1 = 1$ with eigenvalue λ satisfying $\mathrm{Re}\lambda > 0$. Let $h(\cdot)$ satisfy $|h'| \leq C_1 h$ for some constant C_1. If*

$$|\partial_e \mu_\pm(\langle v\rangle \mp \beta, v_2)| \leq C_2 h(\langle v\rangle)\mu_\pm(\langle v\rangle \mp \beta, v_2) \tag{60}$$

and $|\beta|_{C^1}$ is sufficiently small, then

$$|R_\pm(x,v)| \leq C_3 h(\langle v\rangle)\mu_\pm(\langle v\rangle \mp \beta, v_2),$$

where C_3 depends only on C_1, C_2, $\mathrm{Re}\lambda$, and $|\beta'|_\infty$.

Proof. Omit the subscripts $\pm$. The eigenfunction satisfies

$$[\hat{v}_1\partial_x \pm \beta'(x)\partial_{v_1}]R \pm E_0\partial_{v_1}\mu = -\lambda R$$

where $\mu = \mu_\pm(\langle v\rangle \mp \beta'(x), v_2)$. Then $S = (h\mu)^{-1}R$ satisfies

$$[\hat{v}_1\partial_x \pm \beta'(x)\partial_{v_1}]S \pm E_0\frac{\partial_{v_1}\mu}{h\mu} \pm \beta'\frac{\partial_{v_1}h}{h}S = -\lambda S,$$

where we have used the fact that $\hat{v}_1\partial_x\mu_\pm \pm \beta'\partial_{v_1}\mu_\pm = 0$. As in (25), this may be written as

$$S = \mp\int_0^\infty e^{-t\mathcal{A}}e^{-\lambda t}\left[E_0\frac{\partial_{v_1}\mu}{h\mu} + \beta'\frac{\partial_{v_1}h}{h}S\right]dt.$$

Since $\exp(-t\mathcal{A})$ has norm one on L^∞, for $\mathrm{Re}\lambda > 0$ we have

$$\begin{aligned}\|S\|_\infty &\leq \left[|E_0|_\infty\left\|\frac{\partial_{v_1}\mu}{h\mu}\right\|_\infty + |\beta'|_\infty\left\|\frac{\partial_{v_1}h}{h}\right\|_\infty\|S\|_\infty\right](\mathrm{Re}\lambda)^{-1}\\ &\leq [C_2|E_0|_\infty + C_1|\beta'|_\infty\|S\|_\infty](\mathrm{Re}\lambda)^{-1}.\end{aligned}$$

Since $|E_0|_\infty \leq |E_0|_{1,1} \leq C\|R\|_1$, the lemma thus follows if $|\beta'|_\infty$ is small. ∎

The following lemma gives an improved bound for a cut-off eigenfunction.

LEMMA 14 (approximate eigenfunctions). *Let $[R_\pm, E_0]$ and β be as in the preceding lemma. Let $h(s)$ be either s^σ or $\exp(ls)$, for some $\sigma > 0$ or $l > 0$. Assume there are constants C_2, C_5, and $m_0 > 0$ such that for sufficiently large s,*

$$|\partial_e \mu_\pm(s, v_2)| \leq C_2 h(s)\mu_\pm(s, v_2), \tag{61}$$

$$\mu_\pm(s, v_2) \leq C_5 h'(s)[h(s)]^{-(2+m_0)}s^{-1}. \tag{62}$$

Then there exists $\delta_0 > 0$ such that for $0 < \delta < \delta_0$, there exist approximate eigenfunctions $R^\delta_\pm \in L^1(\mathbf{R}_P \times \mathbf{R})$, $E^\delta_0 \in L^1(\mathbf{R}_P)$ such that all of the following hold:

$$\delta |R^\delta_\pm(x,v)| \le \mu_\pm(\langle v\rangle \mp \beta(x), v_2), \tag{63}$$

$$\|R^\delta_\pm - R_\pm\|_1 + |E^\delta_0 - E_0|_1 \le \delta^m, \tag{64}$$

$$\int_0^P \int_{\mathbf{R}} (R^\delta_+ - R^\delta_-) dv dx = 0, \tag{65}$$

$$\partial_x E^\delta_0 = \int_{\mathbf{R}} (R^\delta_+ - R^\delta_-) dv, \tag{66}$$

$$\|R^\delta\|_{1,1} \le C\|R\|_1, \tag{67}$$

where $0 < m < m_0$. Furthermore, there exists a disk Ω^δ independent of x such that $R^\delta(x,v)$ has support in $\mathbf{R} \times \Omega^\delta$.

Proof. We prove this lemma in two steps.

Step 1. *Cut-off approximation.* Let $\eta(\langle v\rangle)$ be a smooth cut-off function, $\eta(\langle v\rangle) = 1$ for $\langle v\rangle \le w$, $\eta(\langle v\rangle) = 0$ for $\langle v\rangle \ge w+1$, with w to be chosen later. Notice that (61) implies (60). By Lemma 13,

$$|\eta(\langle v\rangle) R_\pm(x,v)| \le |R_\pm(x,v)| \le C_3 h(\langle v\rangle) \mu_\pm(\langle v\rangle \mp \beta(x), v_2).$$

Define w by the equation $\delta = [2C_3 h(w+1)]^{-1}$. Then $\delta\eta(v)|R_\pm(x,v)| \le \frac{1}{2}\mu_\pm(\langle v\rangle \mp \beta(x), v_2)$ since h is an increasing function. Now from (62) we have

$$\mu_\pm(s, v_2) = o[h'(s)(h(s))^{-(2+m)}] s^{-1}$$

as $s \to \infty$ uniformly in v_2. Hence

$$\int_0^P \mu_\pm(\langle v\rangle \mp \beta(x), v_2) dx = o\left\{ \frac{h'(\langle v\rangle)}{[h(\langle v\rangle)]^{2+m}\langle v\rangle} \right\}$$

as $\langle v\rangle \to \infty$. Integrating this equality, we get

$$\int_{\langle v\rangle \ge w} \int_0^P h(\langle v\rangle) \mu_\pm(\langle v\rangle \mp \beta(x), v_2) dx dv = o\left\{ \int_{\langle v\rangle \ge w} \int_0^P \frac{h'(\langle v\rangle)}{[h(\langle v\rangle)]^{1+m}\langle v\rangle} dx dv \right\} \le \delta^m$$

for sufficiently small δ, by the definition of w. Hence,

$$\int_{\mathbf{R}^2} \int_0^P |\eta R_\pm - R_\pm| dx dv \le \int_{\langle v\rangle \ge w} \int |R_\pm| dx dv$$

$$\le C \int_{\langle v\rangle \ge w} \int_0^P h(\langle v\rangle) \mu_\pm(\langle v\rangle \mp \beta, v_2) dx dv \le C_6 \delta^m$$

for sufficiently large w. Reducing m slightly eliminates the constant C_6.

Step 2. *Neutrality and Poisson conditions.* We now further perturb the cut-off eigenfunctions. Let $0 \le Q(v) \in C^1_0(\mathbf{R}^2)$, $P\int Q(v) dv = 1$. By Step 1, we define for every $\delta > 0$,

$$R^\delta_+ = \eta R_+ + aQ, \qquad R^\delta_- = \eta R_-$$

where a is a complex number satisfying (65)

$$\int_0^P \int_{\mathbf{R}} (R^\delta_+ - R^\delta_-) dx dv = a + \int_0^P \int_{\mathbf{R}} \eta(R_+ - R_-) dv dx = 0.$$

By Step 1 and the neutrality condition (20),

$$|a| = \left| \int \int \eta(R_+ - R_-)dxdv \right| = \left| \int \int (1-\eta)(R_+ - R_-)dxdv \right| \le \delta^m.$$

We also deduce (63) from Step 1 because δ is sufficiently small and $\mu_+ > 0$. By an easy calculation and the bound on a, $\|R_\pm^\delta\|_{1,1} \le C\delta^m + C\|R_\pm\|_{1,1} \le C'\|R\|_1$. The last inequality follows from Lemma 11 and the normalization. We finally define E^δ to satisfy

$$\partial_x E_0^\delta = \int_{\mathbf{R}^2} (R_+^\delta - R_-^\delta)dv$$

with the same average as E_0: $\int_0^P E_0^\delta dx = \int_0^P E_0 dx$. Hence

$$|E_0^\delta - E_0|_1 \le C \left| \partial_x [E_0^\delta - E_0] \right|_1 \le C\delta^m.$$

We deduce (64) and (67) for small δ and the lemma follows. □

Remark. Conditions (61) and (62) are very general. They allow $\mu_\pm$ to go to zero at a polynomial, exponential, or even super-exponential rate but they exclude $\mu_\pm$ of compact support. An example is $\mu(s) = \exp[-s^\alpha]$ with $\alpha \ge 1$ and $h(s) = s^{\alpha-1}$. Another example is $\mu(s) = \exp[-\exp s]$ and $h(s) = \exp s$.

5. Nonlinear instability for periodic BGK waves. Let us abbreviate

$$f = [f_+, f_-], \qquad \mu_\beta = [\mu_+(\langle v \rangle - \beta, v_2), \mu_-(\langle v \rangle + \beta, v_2)],$$
$$u = [f_+, f_-, E], \quad \nu_\beta = [\mu_+(\langle v \rangle - \beta, v_2), \mu_-(\langle v \rangle + \beta, v_2), \partial_x \beta].$$

We define the norm

$$\|u\|_1 = \int_{\mathbf{R}^2} \int_0^P (|f_+| + |f_-|)dxdv + \int_0^P |E|dx. \tag{68}$$

Our goal is to show that the P-periodic BGK equilibrium ν_β is nonlinearly unstable under $\|\cdot\|_1$ with $P = 2P_\beta$.

LEMMA 15. *Let $\mu_\pm$ satisfy* (11), (12), (14), (61), *and* (62). *Let $[f_+, f_-, E]$ be a BV solution of the nonlinear Vlasov–Maxwell system as in Theorem* 5 *in the appendix. Let $\omega > 0$ and $\|\beta\|_{C^2} < \omega^2$. Let $\delta > 0$. Assume there are positive constants δ, T, b_0, and C_0 such that*

$$\begin{aligned} &\|f(0) - \mu_\beta\|_{1,1} \le b_0\delta, \\ &\|f(t) - \mu_\beta\|_1 \le C_0\delta e^{\omega t} \end{aligned} \tag{69}$$

for $0 \le t \le T$. Then there are positive constants θ and D depending only on b_0, C_0, and ω such that

$$\|\partial_x[f(t) - \mu_\beta]\|_m + \|\partial_{v_1}[f(t) - \mu_\beta]\|_m \le D\delta e^{\omega t} \tag{70}$$

in $0 \le t \le \min\{T, \frac{1}{\omega}\ln(\theta/\delta)\}$. Here m denotes the measure norm.

Proof. We let $L^1 = L^1(\mathbf{R}_P \times \mathbf{R}^2)$ throughout this proof for notational simplicity. Without loss of generality (by a smooth approximation), we may assume $f_\pm \in W^{1,1}$ so that the measure norms in (70) are replaced by L^1 norms. Taking the x derivative of the Vlasov equation, we get

$$(\partial_t + \hat{v}_1\partial_x \pm E\partial_{v_1})(\partial_x f_\pm) \pm \partial_x E \partial_{v_1} f_\pm = 0.$$

Taking the x-derivative of the stationary Vlasov equation for $\mu_\pm$ yields

$$\hat{v}_1\partial_{xx}\mu_\pm \pm \beta_{xx}\partial_{v_1}\mu_\pm \pm \beta_x\partial_{xv_1}\mu_\pm = 0.$$

The difference of these two equations is

$$\begin{aligned}&(\partial_t + \hat{v}_1\partial_x \pm E\partial_{v_1})[\partial_x(f_\pm - \mu_\pm)]\\ &= \mp E_x\partial_{v_1}(f_\pm - \mu_\pm) \mp (\beta_{xx} - E_x)\partial_{v_1}\mu_\pm \pm (E - \beta_x)\partial_{xv_1}\mu_\pm\end{aligned} \tag{71}$$

where $\mu_\pm = \mu_\pm(\langle v\rangle \mp \beta(x), v_2)$. Notice that $\int(|\partial_{v_1}\mu_\pm| + |\partial_x\partial_{v_1}\mu_\pm|)dv$ is bounded in x from (11). Now multiplying (71) by $sgn[\partial_x(f_\pm - \mu_\pm)]$ and integrating over $[0,P]\times\mathbf{R}^2$, we get for some $\epsilon > 0$,

$$\begin{aligned}&\frac{d}{dt}\|\partial_x[f_\pm - \mu_\pm]\|_1\\ &\le |E_x|_\infty\|\partial_{v_1}[f_\pm - \mu_\pm]\|_1 + C(|\beta_{xx} - E_x|_1 + |\beta_x - E|_1)\\ &\le (|\beta_{xx}|_\infty + |E_x - \beta_{xx}|_\infty)\|\partial_{v_1}[f_\pm - \mu_\pm]\|_1 + C(|\beta_{xx} - E_x|_1 + |\beta_x - E|_1)\\ &\le (|\beta_{xx}|_\infty + \|\partial_x[f_\pm - \mu_\pm]\|_1)\|\partial_{v_1}[f_\pm - \mu_\pm]\|_1 + C(\|f_\pm - \mu_\pm\|_1 + |\beta_x - E|_1)\end{aligned} \tag{72}$$

since $E_x - \beta_{xx}$ has average zero.

Similarly, by taking the v_1 derivative of the Vlasov equation, we get

$$\begin{aligned}&\partial_t\partial_{v_1}f_\pm + \hat{v}_1\partial_{v_1}\partial_x f_\pm \pm E\partial_{v_1}\partial_{v_1}f_\pm = -\langle v\rangle^{-3}(1+v_2^2)\partial_x f_\pm,\\ &\hat{v}_1\partial_x\partial_{v_1}\mu_\pm \pm \beta_x\partial_{v_1}\partial_{v_1}\mu_\pm = -\langle v\rangle^{-3}(1+v_2^2)\partial_x\mu_\pm.\end{aligned}$$

Taking the difference yields

$$(\partial_t + \hat{v}_1\partial_x \pm E\partial_{v_1})(\partial_{v_1}[f_\pm - \mu_\pm]) = -\langle v\rangle^{-3}(1+v_2^2)\partial_x[f_\pm - \mu_\pm] \mp (E - \beta_x)\partial_{v_1}\partial_{v_1}\mu_\pm. \tag{73}$$

We also have

$$\|(E - \beta_x)\partial_{v_1}\partial_{v_1}\mu_\pm\|_1 \le C|E - \beta_x|_1.$$

We have used the fact that $\sup_x \int_{\mathbf{R}^2} |\partial_{v_1}\partial_{v_1}\mu_\pm|dv < \infty$ by (11). Multiplying (73) by $sgn(\partial_{v_1}(f_\pm - \mu_\pm))$ and integrating over $[0,P]\times\mathbf{R}$, we get

$$\frac{d}{dt}\|\partial_{v_1}[f_\pm - \mu_\pm]\|_1 \le \|\partial_x[f_\pm - \mu_\pm]\|_1 + C|E - \beta_x|_1.$$

With D to be chosen later larger than b_0, define T' so that $[0,T']$ is the maximal interval in which (70) is valid. Since $\|\beta\|_{C^2} < \omega^2$, we may fix $0 < \epsilon < \omega - \|\beta\|^{1/2}$. Then choose θ so small that $\|\beta\| + C_1\theta < (\omega - \epsilon)^2$ and define T'' by $\delta\exp(\omega T'') = \theta$. Then T'', T, and θ depend on D but the other constants C and C_ϵ (below) do not. Then

$$\|\beta_{xx}\|_\infty + \|\partial_x[f_\pm - \mu_\pm]\|_1 < (\omega - \epsilon)^2$$

for $0 \le t \le \min\{T, T', T''\}$. Therefore, integrating (72) and plugging it into the above inequality, we have for $0 \le t \le \min\{T, T', T''\}$,

$$\begin{aligned}\frac{d}{dt}\|\partial_{v_1}[f_\pm - \mu_\pm]\|_1 \le&(\omega - \epsilon)^2\int_0^t \|\partial_{v_1}[f_\pm(\tau) - \mu_\pm]\|_1 d\tau + \|f_\pm(0) - \mu_\pm\|_{1,1}\\ &+C\int_0^t \{\|f_\pm(\tau) - \mu_\pm\|_1 + |E(\tau) - \beta_x|_1\}d\tau + C|E(t) - \beta_x|_1.\end{aligned}$$

Because of the identity $\partial_t(E-\beta_x)=-j$, we have $\frac{d}{dt}|E-\beta_x|_1 \le \|f-\mu\|_1$, so that $|E(t)-\beta_x|_1 \le C\delta e^{\omega t}$ by (69). Letting $V(t)=\int_0^t \|\partial_{v_1}[f_\pm(\tau)-\mu_\pm]\|_1 d\tau$, we therefore have by (69)

$$V'' \le (\omega-\epsilon)^2 V + \frac{1}{2}C\delta e^{\omega t}$$

for some constant C. Multiplying by $2V'$ on both sides, we get

$$[(V')^2]' \le [(\omega-\epsilon)^2 V^2]' + C\delta e^{\omega t} V'.$$

Integrating over time, we get

$$\begin{aligned}(V')^2 &\le (V'(0))^2 + (\omega-\epsilon)^2V^2 + C\delta\left[e^{\omega t}V - \omega\int_0^t e^{\omega t}V(\tau)d\tau\right]\\ &\le b_0^2\delta^2 + (\omega-\epsilon)^2V^2 + C\delta e^{\omega t}V\\ &\le (\omega-\epsilon/2)^2V^2 + C_\epsilon^2\delta^2 e^{2\omega t}.\end{aligned}$$

Taking the square root of both sides, we obtain $V' \le (\omega-\epsilon/2)V + C_\epsilon\delta e^{\omega t}$. It follows from this inequality and (72) that

$$\|\partial_x[f(t)-\mu]\|_1 + \|\partial_{v_1}[f(t)-\mu]\|_1 \le C'_\epsilon\delta e^{\omega t}$$

in $[0,\min\{T,T',T''\}]$ for some constant C'_ϵ independent of D. We choose $D > C'_\epsilon$. Then $\min\{T,T''\} \le T'$ and the lemma follows. □

We now are ready to prove the nonlinear instability of periodic BGK waves.

THEOREM 3. *Let $\mu_\pm$ satisfy* (11), (12), (14), (61), *and* (62). *Let β be a solution of* (4) *of period P_β with $\|\beta\|_{C^2} \le \beta_0$, where β_0 is sufficiently small. Consider $1\frac{1}{4}$RVM with $P=2P_\beta$. Then there exist positive constants ϵ_0 and C_1 and a family of solutions $u^\delta(t)=[f^\delta_\pm(t),E^\delta_1(t)]$ of $1\frac{1}{4}$ RVM with $f^\delta_\pm \ge 0$ defined for δ sufficiently small, such that*

$$\sum_\pm \|f^\delta_\pm(0)-\mu_\pm(\langle v\rangle \mp \beta, v_2)\|_{W^{1,1}(\mathbf{R}_P\times\mathbf{R}^2)} + |E^\delta_1(0)-\partial_x\beta|_{W^{1,1}(\mathbf{R}_P)} \le \delta$$

and

$$\sup_{0\le t\le C_1|\ln\delta|} \|u^\delta(t)-\nu_\beta\|_1 \ge \epsilon_0.$$

Proof of Theorem 3. We are given nonnegative $\mu_\pm$ that satisfy (11), (12), (14), (61), and (62). Furthermore, β is a solution of (4) of period P_β with $\|\beta\|_{C^2}$ sufficiently small as in Lemma 1 and $\nu_\beta=[\mu_\beta,\beta_x]$. We must find a family of solutions $u^\delta(t)=[f^\delta_\pm(t),E^\delta(t)]$ of the nonlinear Vlasov–Maxwell system satisfying the conclusions of Theorem 5, such that

$$\begin{aligned}&\|f^\delta(0)-\mu_\beta\|_{1,1} + |E^\delta(0)-\beta_x|_{1,1} \le \delta,\\ &\sup_{0\le t<C_1|ln\delta|}\|u^\delta(t)-\nu_\beta\|_1 \ge \epsilon_0 > 0\end{aligned} \tag{74}$$

with $\|\cdot\|_1$ defined by (68).

By Lemma 1, the BGK equilibria exist because of (11), (12), and (14). By Theorem 2 and Lemma 11 and because of (11) and (12), we may choose $\Xi=[R_+,R_-,E_0]$

to be an eigenvector of $-\mathcal{L}$ satisfying (29) with $\|R\|_{1,1}+|E_0|_1=1$ such that its eigenvalue λ has the largest (positive) real part. If λ is not real, then $\|\mathrm{Im}\ \Xi\|_1 \equiv r>0$ by Lemma 12. We choose an approximation $\Xi^\delta=[\mathrm{Im}\ R^\delta_\pm, \mathrm{Im}\ E^\delta_0]$ to the imaginary part of Ξ by Lemma 14. In case λ is real we simply do not take the imaginary parts; but without loss of generality, we will assume λ is not real.

We choose the family of solutions $u^\delta(t,x,v)=[f^\delta_\pm(t,x,v), E^\delta(t,x)]$ by specifying the initial data $u^\delta(0,x,v)=\nu_\beta+\delta\Xi^\delta$. That is,

$$f^\delta_\pm(0,x,v)=\mu_\pm(\langle v\rangle \mp \beta(x), v_2)+\delta\mathrm{Im}\ R^\delta_\pm(x,v), E^\delta(0,x)=\beta_x(x)+\delta\mathrm{Im}\ E^\delta_0(x).$$

Because of (63), $f^\delta_\pm(0,x,v)\ge 0$ for all x, v and for all sufficiently small δ. Because of Lemma 14, all of the conditions of Theorem 5 are satisfied. Note that

$$|\|u(0)-\nu_\beta\|_1-\delta r| = \delta(\|\Xi^\delta\|_1-r)\le \delta\|\Xi-\Xi^\delta\|_1\le \delta^{m+1}\le \delta r/2 \tag{75}$$

by (64) for δ sufficiently small. By (67)

$$\begin{aligned}\|f(0)-\mu_\beta\|_{1,1}+|E(0)-\beta_x|_1 &= \delta\|\mathrm{Im}\ R^\delta\|_{1,1}+\delta|\mathrm{Im}\ E^\delta_0|_1\\ &\le C\delta[\|\mathrm{Im}\ R^\delta\|_1+|\mathrm{Im}\ E^\delta_0|_1]=C\delta r.\end{aligned} \tag{76}$$

Let $u^\delta(t)=u(t)=[f_+(t),f_-(t),E(t)]$ denote the solution, where we drop the superscript δ.

By the nonlinear Vlasov–Maxwell system

$$u(t)-\nu_\beta=\delta e^{-\mathcal{L}t}\Xi^\delta+\int_0^t e^{-\mathcal{L}(t-\tau)}\begin{pmatrix}\mp(E(\tau)-\beta_x)\partial_{v_1}(f_\pm(\tau)-\mu_\pm)\\ 0\end{pmatrix}d\tau. \tag{77}$$

We choose Λ such that

$$\mathrm{Re}\lambda<\Lambda<\min(1+m,2)\mathrm{Re}\lambda. \tag{78}$$

Let $\|\beta\|<(\mathrm{Re}\lambda)^2=\omega^2$. We define S by

$$\delta e^{\mathrm{Re}\lambda S}=\{\zeta r/(2C_\Lambda)\}^{1/m}.$$

Let C_Λ be the constant in Lemma 10 and ζ be the constant in Lemma 12. Let

$$T=\sup\left\{s:\ \|u(t)-\nu_\beta-\delta e^{-\mathcal{L}t}\Xi^\delta\|_1\le\frac{\zeta}{4}\,\delta e^{\mathrm{Re}\lambda t}r,\ \text{for } 0\le t\le s\right\}. \tag{79}$$

For $0\le t\le\min\{S,T\}$, from Lemma 10 and (64) and (78),

$$\|e^{-\mathcal{L}t}(\mathrm{Im}\ \Xi-\Xi^\delta)\|_1\le C_\Lambda e^{\Lambda t}\delta^m\le C_\Lambda e^{\omega t}\{\delta e^{\omega t}\}^m\le\frac{1}{2}\zeta r e^{\omega t}$$

by choice of S. Hence, by (79) for such t,

$$\begin{aligned}\|u(t)-\nu_\beta\|_1&\le\delta e^{\mathrm{Re}\lambda t}\|\Xi\|_1+\delta\|e^{-\mathcal{L}t}(\mathrm{Im}\ \Xi-\Xi^\delta)\|_1+\frac{\zeta}{4}\delta e^{\mathrm{Re}\lambda t}r\\ &\le(1+3\zeta r/4)\delta e^{\mathrm{Re}\lambda t}.\end{aligned} \tag{80}$$

Hence for such t, from

$$\partial_x(E-\beta_x)=\rho-\beta_{xx}=\int_{\mathbf{R}^2}[(f_+-\mu_+)-(f_--\mu_-)]dv,$$

we deduce

$$
\begin{aligned}
|E(t)-\beta_x|_\infty \le & \frac{1}{P}|E(t)-\beta_x|_1+\sum_{\pm}\|f_\pm(t)-\mu_\pm\|_1 \\
\le & C\|u(t)-\nu_\beta\|_1 \le C\delta e^{t\mathrm{Re}\lambda}. \qquad (81)
\end{aligned}
$$

By Lemma 15, there exist D and $\theta>0$ such that if $\delta e^{\mathrm{Re}T^*}=\theta$ and $0\le t\le \min\{T,S,T^*\}$, then

$$
\|\partial_{v_1}[f(t)-\mu_\beta]\|_m \le D\delta e^{\mathrm{Re}\lambda t}.
$$

We may assume $\theta \le \{\zeta r/(2C_\Lambda)\}^{1/m}$ so that $T^*\le S$. Hence for such t, by Lemma 10 and (77) and (81),

$$
\begin{aligned}
&\|u(t)-\nu_\beta-\delta e^{-\mathcal{L}t}\Xi^\delta\|_1 \\
&\le C\int_0^t e^{\Lambda(t-\tau)}|E(\tau)-\partial_x\beta|_\infty \sum_{\pm}\|\partial_{v_1}[f_\pm(\tau)-\mu_\pm]\|_m d\tau \\
&\le C\int_0^t e^{\Lambda(t-\tau)}(\delta e^{\tau\mathrm{Re}\lambda})^2 d\tau \le C_2(\delta e^{\mathrm{Re}\lambda t})^2 \qquad (82)
\end{aligned}
$$

since $\Lambda<2\mathrm{Re}\lambda$, with a constant C_2 independent of θ, δ, and t. Thus for $0\le t\le \min\{T,T^*\}$, we also have

$$
\begin{aligned}
\|u(t)-\nu_\beta\|_1 &\ge \delta\|e^{-\mathcal{L}t}\Xi^\delta\|_1-\|u(t)-\nu_\beta-\delta e^{-\mathcal{L}t}\Xi^\delta\|_1 \\
&\ge \delta\|e^{-\mathcal{L}t}\mathrm{Im}\Xi\|_1-\delta\|e^{-\mathcal{L}t}(\mathrm{Im}\Xi-\Xi^\delta)\|_1-C_2(\delta e^{\mathrm{Re}\lambda t})^2 \qquad (83) \\
&\ge \frac{1}{2}\delta r\zeta e^{\mathrm{Re}\lambda t}-C_2(\delta e^{\mathrm{Re}\lambda t})^2
\end{aligned}
$$

by Lemma 12 and as in (80).

If $T<T^*$, then by (82) with $t=T$, we have

$$
\begin{aligned}
\|u(T)-\nu_\beta-\delta e^{-\mathcal{L}T}\Xi^\delta\|_1 &\le C_2(\delta e^{\mathrm{Re}\lambda T})^2 \\
&< C_2(\delta e^{\mathrm{Re}\lambda T})\theta < \frac{\zeta}{4}r\delta e^{T\mathrm{Re}\lambda}
\end{aligned}
$$

by also choosing $0<\theta<\frac{\zeta r}{4C_2}$. This contradicts (79). Therefore $T^*\le T$. By (83) with $t=T^*$, we have

$$
\|u(T^*)-\nu_\beta\|_1 \ge \frac{r\zeta}{2}\delta e^{\omega T^*}-C_2(\delta e^{\omega T^*})^2=\frac{r\zeta}{2}\theta-C_2\theta^2>\frac{r\zeta}{4}\theta
$$

since $0<\theta<\frac{r\zeta}{4C_2}$. □

6. Instability of oscillatory-tail solutions. In this section, we study the instabilities of oscillatory-tail solutions of (1). A major difficulty lies in the unboundedness of the spatial variable, so that the plane wave growing mode does not decay as $x\to\infty$. They do not belong in any L^p space, and they correspond to continuous spectrum. We shall overcome this by employing the finite propagation speed property of the relativistic model. We approximate the original problem on the whole line by the asymptotic periodic problem.

Consider a plasma which consists of electrons and ions in one space dimension with orthogonal electric and magnetic fields $(E_1, E_2, 0)$ and $(0, 0, B)$. By normalizing all the physical constants to be one, we derive the $1\frac{1}{2}$-dimensional RVM system as

$$\begin{aligned} &\partial_t f_\pm + \hat{v}_1 \partial_x f_\pm \pm [E_1 + \hat{v}_2 B]\partial_{v_1} f_\pm \pm [E_2 - \hat{v}_1 B]\partial_{v_2} f_\pm = 0, \\ &\partial_t E_1 = -j_1, \qquad \partial_x E_1 = \rho, \\ &\partial_t E_2 = -\partial_x B - j_2, \qquad \partial_t B = -\partial_x E_2. \end{aligned} \tag{84}$$

Here $f_\pm(t, x, v_1, v_2)$ are the microscopic distribution functions for ions $(+)$ and electrons $(-)$ at time t, position x, and momentum (v_1, v_2). The relativistic velocity is $\hat{v} = \frac{v}{\sqrt{1+|v|^2}}$. The charge and current densities are defined by

$$\rho = \int_{\mathbf{R}^2} [f_+ - f_-] dv_1 dv_2, \qquad j_k = \int_{\mathbf{R}^2} \hat{v}_k [f_+ - f_-] dv_1 dv_2$$

for $k = 1, 2$.

For the well-posedness of (84) see Theorem 4 in the appendix. For any interval $A \subset \mathbf{R}$, define the norm

$$\|u\|_A = \sum_\pm \|f_\pm\|_{1,1} + |E|_1 + |E_2|_1 + |B|_1,$$

where the x-norms are taken over the interval A. Recalling Γ and Γ_0 in (6) and (5), we have a lemma.

LEMMA 16. *Let I be the interval $I = [a - b, a + b]$. Then for all b,*

$$\lim_{a \to -\infty} \|\Gamma - \Gamma_0\|_I = 0.$$

Proof of Lemma 16. We write $\|\Gamma - \Gamma_0\|_{L^1(I)} = I_1 + I_2 + I_3$ where we abbreviate $\mu_\pm = \mu_\pm(\langle v \rangle \mp \Phi(x), v_2 \pm \Psi(x))$,

$$\begin{aligned} I_1 &= \sum_\pm \int_{\mathbf{R}^2} \int_I |\mu_\pm - \mu_\pm(\langle v \rangle \mp \beta(x), v_2)| \, dx dv, \\ I_2 &= \int_I \{|\partial_x(\Phi - \beta)| + |\partial_x \Psi|\} dx, \\ I_3 &= \sup_{x \in I} \left| \int_{\mathbf{R}^2} \{\mu_+ - \mu_+(\langle v \rangle - \beta(x), v_2) - \mu_- + \mu_-(\langle v \rangle + \beta(x), v_2)\} dv \right|. \end{aligned}$$

By (7) it follows immediately that $\lim_{a \to -\infty} I_2 = 0$. By the decay assumption on $\mu_\pm$ in (11),

$$\begin{aligned} \Delta &= |\mu_\pm(\langle v \rangle \mp \Phi(x), v_2 \pm \Psi(x)) - \mu_\pm(\langle v \rangle \mp \beta(x), v_2)| \\ &\le C \langle \langle v \rangle \mp \beta(x) \pm \theta_1 \rangle^{-\gamma} \langle v_2 \pm \theta_2 \rangle^{-\tilde{\gamma}} |\Phi(x) - \beta(x)| \\ &\quad + C \langle \langle v \rangle \mp \beta(x) \pm \theta_1 \rangle^{-\gamma} \langle v_2 \pm \theta_2 \rangle^{-\tilde{\gamma}} |\Psi(x)|, \end{aligned}$$

where θ_1 lies between 0 and $\Phi(x) - \beta(x)$, and θ_2 between 0 and $\Psi(x)$. Hence

$$\Delta \le C \langle v \rangle^{-\gamma} \langle v_2 \rangle^{-\tilde{\gamma}} (|\Phi(x) - \beta(x)| + |\Psi(x)|).$$

By assumption (11), these factors of v are integrable over $\mathbf{R}^2$. Hence

$$\int_{\mathbf{R}^2} \Delta dv \le C|\Phi(x) - \beta(x)| + C|\Psi(x)| \tag{85}$$

so that, by (7), $\lim_{a\to-\infty}(I_1+I_3)=0$.

Now we write $\|\|\Gamma-\Gamma_0\|\|_I=\|\Gamma-\Gamma_0\|_{L^1(I)}+I_4+I_5$, where

$$I_4=\sum_{\pm}\int_{\mathbf{R}^2}\int_I\sum_{|\sigma|=1}|\partial^\sigma\{\mu_\pm-\mu_\pm(\langle v\rangle\mp\beta(x),v_2)\}|\,dxdv,$$

$$I_5=\int_I\{|\partial_x(\Phi-\beta)|+|\partial_x\Psi|\}dx.$$

Now

$$I_5\le 2b\sup_I\{|\partial_x(\Phi-\beta)|+|\partial_x\Psi|\}\to 0,$$

provided a is chosen sufficiently near $-\infty$. This takes care of I_5. Now I_4 consists of several terms. The first one is estimated as

$$\begin{aligned}|\partial_{v_1}\{\mu_\pm-\mu_\pm(\langle v\rangle\mp\beta(x),v_2)\}|&\le|\partial_e\mu_\pm-(\partial_e\mu_\pm)(\langle v\rangle\mp\beta(x),v_2)|\\&\le C\langle v\rangle^{-\gamma}\langle v_2\rangle^{-\tilde\gamma}(|\Phi-\beta|+|\Psi|)\end{aligned}$$

in the same way as we treated Δ above. Similarly, by (11), the second term is

$$\begin{aligned}|\partial_{v_2}\{\mu_\pm-\mu_\pm(\langle v\rangle\mp\beta(x),v_2)\}|&\le\sup(|\partial_e^2\mu|+|\partial_e\partial_{v_2}\mu|+|\partial_{v_2}^2\mu|)(|\Phi-\beta|+|\Psi|)\\&\le C\langle v\rangle^{-\gamma}\langle v_2\rangle^{-\tilde\gamma}(|\Phi-\beta|+|\Psi|).\end{aligned}$$

The third term is

$$\begin{aligned}&|\partial_x\{\mu_\pm-\mu_\pm(\langle v\rangle\mp\beta(x),v_2)\}|\\&=|\mp\Phi'(x)\partial_e\mu_\pm\pm\Psi'(x)\partial_{v_2}\mu_\pm\pm\beta'(x)\partial_e\mu_\pm(\langle v\rangle\mp\beta(x),v_2)|\\&\le|\Phi'-\beta'|\sup|\partial_e\mu_\pm|+|\Psi'|\sup|\partial_{v_2}\mu_\pm|\\&+|\beta'||\Phi-\beta|\sup|\partial_e^2\mu_\pm|+|\beta'||\Psi'|\sup|\partial_e\partial_{v_2}\mu_\pm|\\&\le C\langle v\rangle^{-\gamma}\langle v_2\rangle^{-\tilde\gamma}\{|\Phi'-\beta'|+|\Psi'|+|\Phi-\beta|+|\Psi|\}.\end{aligned}$$

These terms are treated in the same way as (85) to obtain

$$I_4\le\int_I\{|\Phi'-\beta'|+|\Psi'|+|\Phi-\beta|+|\Psi|\}dx\to 0$$

provided $a\to-\infty$. This proves the lemma. □

Proof of the main theorem. We will break the x-axis into certain intervals. Let N be a positive integer and $\delta=\exp[-NP/C_1]$, where C_1 will be chosen later. We also choose a number a near $-\infty$ and define intervals

$$I=\{x:\ |x-a|\le(N+2)P\},\qquad J=\{x:\ |x-a|\le(N+1)P\},$$

and $K=\{x:\ |x-a|\le P/2\}$. In Theorem 3 on periodic equilibria, we constructed, for some $\epsilon_0>0$, a family of periodic initial data

$$u_P^\delta(0,x,v)=[f_{P+}^\delta,f_{P-}^\delta,E_{P1}^\delta,0,0]$$

with $f_{P\pm}^\delta\ge 0$, $E_2^\delta=0, B^\delta=0$. This family is defined for all sufficiently small $\delta>0$ and satisfies

$$\epsilon_0<\sup_{0\le t\le C_1|\ln\delta|}\|u_P^\delta(t)-\Gamma_0\|_{L^1[0,P]}\tag{86}$$

but

$$\|u_P^\delta(0) - \Gamma_0\|_{W^{1,1}[0,P]} < \delta. \tag{87}$$

We now define *nonperiodic* initial data for (84) as

$$u^\delta(0) = u_P^\delta(0) \qquad \text{for } x \in J, \qquad u^\delta(0) = \Gamma \qquad \text{for } x \notin I. \tag{88}$$

In the transition zones $L^\pm = \{(N+1)P < \pm(x-a) < (N+2)P\}$, we define the initial data $u^\delta(0) = [f_\pm^\delta(0), E_1^\delta(0), 0, B^\delta(0)]$ as follows. It is consistent with (88) to define $E_2^\delta(0,x) \equiv 0$. We define $B^\delta(0,x)$ as the linear interpolate between 0 and $\partial_x \Psi$. We define $\overline{f}_\pm^\delta(0,x,v)$ as the linear interpolate between $f_{\pm P}^\delta(0)$ and $\mu_\pm(\langle v\rangle \mp \Phi(x), v_2 \pm \Psi(x))$. By Theorem 3 and the decay of $\mu_\pm$ as in (11),

$$\begin{aligned} &\|\overline{f}_\pm^\delta(0,x,v) - \mu_\pm(\langle v\rangle \mp \beta(x), v_2)\|_{W^{1,1}(L)} + |B^\delta(0,x)|_{L^1(L)} \\ &\le C\delta + C\|\Gamma - \Gamma_0\|_L. \end{aligned} \tag{89}$$

We then define for $x \in L^+$

$$f_\pm^\delta(0,x,v) = \bar{f}_\pm^\delta(0,x,v) + \alpha_\pm Q(x,v),$$

where $0 \le Q \in C_c^\infty(L^+ \times \mathbf{R}^2)$ with $\int_{L^+}\int_{\mathbf{R}^2} Q = 1$. The constants $\alpha_\pm \ge 0$ are chosen so that

$$\int_{L^+}\int_{\mathbf{R}^2} [f_+^\delta(0) - f_-^\delta(0)]dvdx = \Phi'(a+(N+2)P) - E_{1P}^\delta(0, a+(N+1)P). \tag{90}$$

This requires

$$\alpha_+ - \alpha_- = -\int_{L^+}\int_{\mathbf{R}^2} \left[\overline{f}_+^\delta(0) - \overline{f}_-^\delta(0)\right] dvdx + \Phi'(a+(N+2)P) - E_{1P}^\delta(0, a+(N+1)P)$$

so that by (89)

$$\begin{aligned} |\alpha_+ - \alpha_-| \le{}& C\delta + C\|\Gamma - \Gamma_0\|_L + \left|\int_{L^+}\int_{\mathbf{R}^2} [\mu_+(\langle v\rangle - \beta, v_2) - \mu_-(\langle v\rangle + \beta, v_2)]dvdx\right| \\ &+ |\Phi'(b+P) - \beta'(b+P)| + |\beta'(b) - E_{1P}^\delta(0,b)|, \end{aligned}$$

where $b = a + (N+1)P$. The integral vanishes because of the periodicity of $\beta(x)$. By (87) and (7) we deduce

$$|\alpha_+ - \alpha_-| \le C\delta + C\|\Gamma - \Gamma_0\|_L. \tag{91}$$

Therefore from (89) and (91) we have

$$\|f_\pm^\delta(0) - \mu_\pm(\langle v\rangle \mp \beta, v_2)\|_{W^{1,1}(L^+)} \le C\delta + C\|\Gamma - \Gamma_0\|_L. \tag{92}$$

We then define $E_1^\delta(0,x)$ in L^+ as

$$E_1^\delta(0,x) = E_{1P}^\delta(0, a+(N+1)P) + \int_{a+(N+1)P}^x \int_{\mathbf{R}^2} (f_+^\delta(0,y,v) - f_-^\delta(0,y,v))dydv.$$

It follows that $E_1^\delta(0,x)$ is continuous at $a+(N+2)P$ and that $\partial_x E_1^\delta(0,x)=\rho^\delta(0,x)$ for $x\in L^+$. Furthermore,

$$E_1^\delta(0,x)-\partial_x\beta(x)=E_{1P}^\delta(0,b)-\partial_x\beta(b)$$
$$+\int_b^x\int_{\mathbf{R}^2}\left[f_+^\delta(0)-\mu_+(\langle v\rangle-\beta,v_2)-f_-^\delta(0)+\mu_-(\langle v\rangle+\beta,v_2)\right]dydv$$

so that by (87) we have

$$|E_1^\delta(0)-\partial_x\beta|_{L^1(L^+)}\le C\delta. \tag{93}$$

We define these functions on L^- by the same method. Therefore, by (89), (92), and (93),

$$\|u^\delta(0)-\Gamma_0\|_L\le C\delta+C\|\Gamma-\Gamma_0\|_L$$

where C is independent of N and δ.

Then we have

$$\|u^\delta(0)-\Gamma\|_{\mathbf{R}}=\|u^\delta(0)-\Gamma_0\|_J+\|u^\delta(0)-\Gamma_0\|_L+\|\Gamma-\Gamma_0\|_I.$$

By definition and by (87),

$$\|u^\delta(0)-\Gamma_0\|_J=\|u_P^\delta(0)-\Gamma_0\|_J<C(N+1)\delta\le C\delta|\ln\delta|.$$

From the three preceding inequalities and Lemma 16, it follows that for a sufficiently near $-\infty$,

$$\|u^\delta(0)-\Gamma\|_{\mathbf{R}}\le C\delta|\ln\delta|.$$

We claim that

$$\|u^\delta(0)-\Gamma\|_{W^{1,1}(\mathbf{R})}<C\delta|\ln\delta|. \tag{94}$$

Proof of the claim (94). It suffices to prove that

$$|\partial_x[E_1^\delta(0)-\partial_x\Phi]|_{L^1}+|\partial_x[E_2^\delta(0)]|_{L^1}+|\partial_x[B^\delta(0)-\partial_x\Psi]|_{L^1}<C\delta|\ln\delta|.$$

Now $E_2^\delta(0,x)\equiv 0$. Next, $\partial_x E_1^\delta(0)=\rho^\delta(0)$ so that

$$|\partial_x[E_1^\delta(0)-\partial_x\Phi]|_{L^1}=\int[f_+^\delta(0)-\mu_+-f_-^\delta(0)+\mu_-]dv,$$

where $\mu_\pm=\mu_\pm(\langle v\rangle\mp\Phi(x),v_2\pm\Psi(x))$. Hence

$$|\partial_x[E_1^\delta(0)-\partial_x\Phi]|_{L^1}\le\sum_\pm|f_\pm^\delta-\mu_\pm|_1<C\delta|\ln\delta|.$$

Finally, $\partial_x[B^\delta(0)-\partial_x\Psi]=-\partial_x^2\Psi$ in J, is equal to 0 outside I, and is equal to $\pm P^{-1}\partial_x\Psi(a\pm(N+1)P)$ in $L^\pm$. Then

$$|\partial_x[B^\delta(0)-\partial_x\Psi]|_{L^1}\le\int_J|\partial_x^2\Psi|dx+\sum_\pm|\partial_x\Psi(a\pm(N+1)P)|<C\delta|\ln\delta|$$

by choosing a sufficiently near $-\infty$. This proves (94).

We apply the existence theorem in the appendix to these initial data. By *causality* and because $NP = C_1|\ln\delta|$, we have on $K = [a - P/2, a + P/2]$ the inequality

$$\sup_{0\le t\le NP} \|u^\delta(t) - \Gamma_0\|_{L^1(K)} = \sup_{0\le t\le C_1|\ln\delta|} \|u_P^\delta(t) - \Gamma_0\|_{L^1(K)} > \epsilon_0$$

by (86). The instability of Γ follows, since

$$\sup_{0\le t\le C_1|\ln\delta|} \|u^\delta(t) - \Gamma\|_{L^1(K)} > \epsilon_0 - \|\Gamma - \Gamma_0\|_{L^1(K)} \ge \epsilon_0/2$$

by Lemma 16 for a sufficiently near $-\infty$. Now let $\delta' = \delta|\ln\delta|$. Then $|\ln\delta'| > \frac{1}{2}|\ln\delta|$ so that

$$\sup_{0\le t\le 2C_1|\ln\delta'|} \|u_P^\delta(t) - \Gamma_0\|_{L^1(K)} > \epsilon_0.$$

This proves Theorem 1 with δ replaced by δ' and C_1 replaced by $2C_1$. ∎

7. Appendix. In this appendix we present the well-posedness theorems required in the body of the paper. We begin with the full $1\frac{1}{2}$ RVM in (84) on the whole line.

THEOREM 4. *Let $f_\pm^0 \in BV(\Omega_x \times \mathbf{R}^2)$, $f_\pm^0 \ge 0$, $\langle v\rangle^l f_\pm^0 \in L^\infty(\Omega_x \times \mathbf{R}^2)$ for $l > 3$, E^0, $B^0 \in W^{1,\infty}(\Omega_x)$, for every bounded open set $\Omega_x \in \mathbf{R}$ and $\partial_x E_1^0 = \int(f_+^0 - f_-^0)dv$. Then there exists a unique solution $[f_+, f_-, E_1, E_2, B]$ to (84) with initial data $[f_+^0, f_-^0, E_1^0, E_2^0, B^0]$ such that, for any bounded open sets $\Omega_x \in \mathbf{R}$ and $\Omega_t \in \mathbf{R}$, $0 \le f_\pm \in L^\infty(\Omega_t; BV(\Omega_x \times \mathbf{R}^2))$, $\langle v\rangle^l f_\pm \in L^\infty(\Omega_t; L^\infty(\Omega_x \times \mathbf{R}^2))$, E and $B \in L^\infty(\Omega_t; W^{1,\infty}(\Omega_x))$.*

Remark. Greater regularity can be obtained with extra assumptions on the initial data.

Proof. By the method of [GSc], there is a unique global solution provided the initial data are smooth with compact support. So for existence we require only appropriate a priori bounds for *smooth* solutions. Notice that from the causality principle, we need only to estimate solutions locally in space. By the local energy identity over the dependent region of $[x - A, x + A]$, at time t, we have

$$\begin{aligned}
&\sum_\pm \int_{x-A}^{x+A}\int_{R^2} \langle v\rangle f_\pm(t,y,v)dvdy + \frac{1}{2}\int_{x-A}^{x+A}[|E(t,y)|^2 + B(t,y)^2]dy \\
&+ \sum_\pm \int_0^t \int_{R^2} (\langle v\rangle \mp v_1)[f_+ + f_-](\tau, x \mp A \mp [t-\tau], v)dvd\tau \qquad (95)\\
&\le \sum_\pm \int_{x-A-t}^{x+A+t}\int_{R^2} \langle v\rangle f_\pm(0,y,v)dydv + \frac{1}{2}\int_{x-A-t}^{x+A+t}[|E(0,y)|^2 + B(0,y)^2]dy.
\end{aligned}$$

It then follows that $f_\pm \in L^\infty(\Omega_t; L^1(\Omega_x \times \mathbf{R}^2))$ for every Ω_t and Ω_x, and $E, B \in L^\infty_{loc}(L^2_{loc})$. Moreover, $\rho \in L^\infty_{loc}(L^1_{loc})$ so that $E_1 \in L^\infty_{loc}(W^{1,1}_{loc})$ from $\partial_x E_1 = \rho$.

Next we employ the representation formulas for E_2 and B (equations (13) and (14) of [GSc]):

$$\begin{aligned}
E_2(t,x) =& \frac{1}{2}[E_2(0,x-t) + E_2(0,x+t) + B(0,x-t) - B(0,x+t)] \\
&- \int_0^t [j_2(\tau, x-t+\tau) + j_2(\tau, x+t-\tau)]d\tau, \qquad (96)
\end{aligned}$$

$$B(t,x) = \frac{1}{2}[E_2(0,x-t) - E_2(0,x+t) + B(0,x-t) + B(0,x+t)] \\ - \int_0^t [j_2(\tau, x-t+\tau) - j_2(\tau, x+t-\tau)]d\tau.$$

Since $\langle v\rangle \pm v_1 \geq |\hat{v}_2|$, it follows from (95) with $A = 0$ that

$$\left| \int_0^t j_2(\tau, x \mp (t-\tau))d\tau \right| \leq C \int_0^t \int_{\mathbf{R}^2} (\langle v\rangle \mp v_1)[f_+ + f_-](\tau, x \mp (t-\tau), v)dvd\tau$$

is bounded locally in x and t. Therefore (96) implies that $E_2(t,x)$ and $B(t,x)$ are pointwise bounded locally in x and t.

To obtain the weighted estimates for $f_\pm$, we multiply the Vlasov equations by $\langle v\rangle^l$ to get

$$\{\partial_t + \hat{v}_1\partial_x \pm (E + \hat{v}\times B)\cdot\nabla_v\}(\langle v\rangle^l f_\pm) = \pm l\langle v\rangle^{l-1}\hat{v}\cdot E f_\pm$$

where $E + \hat{v}\times B = [E_1 + \hat{v}_2 B, E_2 - \hat{v}_1 B, 0]$, $\nabla_v f_\pm = [\partial_{v_1} f_\pm, \partial_{v_2} f_\pm, 0]$, and we have used the fact that $\hat{v}\times B\cdot\nabla_v(\langle v\rangle^l) = 0$. By the standard L^∞ estimate for $\langle v\rangle^l f_\pm$ along the backward trajectory $\frac{dx}{dt} = \hat{v}_1$, $\frac{dv}{dt} = \pm[E + \hat{v}\times B]$, we have

$$\sup_{\mathbf{R}^2}\{\langle v\rangle^l f_\pm(t,x,v)\} \leq \sup_{[x-t,x+t]\times\mathbf{R}^2}\{\langle v\rangle^l f_\pm(0,y,v)\} \\ + l\int_0^t \sup_{[x-(t-\tau),x+(t-\tau)]\times\mathbf{R}^2}\{|E(\tau,y)|\langle v\rangle^l f_\pm(\tau,y,v)\}d\tau.$$

Since E is bounded on bounded sets, we obtain

$$\langle v\rangle^l f_\pm(t) \in L^\infty(\Omega_t\times\Omega_x\times\mathbf{R}^2)$$

for all Ω_t and Ω_x. Since $l > 2$, we deduce

$$|\rho(t,x)| = \left|\int_{\mathbf{R}^2}(f_+ - f_-)dv\right| \leq \sum_\pm \int \sup_{\mathbf{R}^2}\left\{\langle v\rangle^l |f_\pm(t,x,v)|\right\}\langle v\rangle^{-l}dv$$

is also bounded on bounded sets. Since $\partial_x E_1 = \rho$, it follows that $E_1 \in L^\infty_{loc}(W^{1,\infty}_{loc})$.

In order to estimate the derivatives of E_2 and B, it suffices by (96) to estimate $\partial_x\int_0^t j_2(\tau, x\mp t\pm\tau)d\tau$. To this end, we use the splitting method of Lemma 3 of [GSc]. Define $T_\pm = \partial_t \pm \partial_x$ and $S = \partial_t + \hat{v}_1\partial_x$. From Lemma 3 of [GSc], we obtain

$$\partial_x\int_0^t j_2(\tau, x-t+\tau)d\tau = \int_0^t\int \frac{\hat{v}_2}{1-\hat{v}_1}(T_+ - S)[(f_+ - f_-)(\tau, x-t+\tau, v)]dvd\tau \\ = I^+ + I^-.$$

Notice that $T_+ f_+(\tau, x-t+\tau, v) = \frac{d}{d\tau}f_+(\tau, x-t+\tau, v)$ and $Sf_+ = -\nabla_v\{(E+\hat{v}\times B)f_+\}$. By integrating along a side of the dependent triangle, we estimate I^+ as

$$I^+ = \int_0^t \frac{d}{d\tau}\int \frac{\hat{v}_2}{1-\hat{v}_1}f_+(\tau, x-t+\tau, v)dvd\tau \\ + \int_0^t\int \frac{\hat{v}_2}{1-\hat{v}_1}\nabla_v([E+\hat{v}\times B]f_+)(\tau, x-t+\tau, v)dvd\tau \\ = \int \frac{\hat{v}_2}{1-\hat{v}_1}f_+(t,x,v)dv - \int \frac{\hat{v}_2}{1-\hat{v}_1}f_+(0, x-t, v)dv \\ - \int_0^t\int \nabla_v\left(\frac{\hat{v}_2}{1-\hat{v}_1}\right)([E+\hat{v}\times B]f_+)(\tau, x-t+\tau, v)dvd\tau.$$

Since $|\frac{\hat{v}_2}{1-\hat{v}_1})| \le \langle v\rangle$, we get $|\frac{\hat{v}_2}{1-\hat{v}_1} f_+(t,x,v)| \le C\langle v\rangle^{-l+1}$ on bounded sets, thereby bounding the first term in I^+. Notice that

$$\partial_{v_1}\left(\frac{\hat{v}_2}{1-\hat{v}_1}\right) = \frac{v_2}{\langle v\rangle(\langle v\rangle - v_1)} = O(1)$$

and

$$\partial_{v_2}\left(\frac{\hat{v}_2}{1-\hat{v}_1}\right) = -\frac{1}{\langle v\rangle}\left[\frac{v_1}{\langle v\rangle - v_1} - \frac{1}{(\langle v\rangle - v_1)^2}\right] = O(\langle v\rangle)$$

since $\langle v\rangle - v_1 \ge \frac{1}{2\langle v\rangle}$. Since E and B are also bounded locally and $\langle v\rangle^l f_\pm$ is bounded locally in x and in t, it follows that the last term in I^+ is bounded by

$$C\int_0^t \int_{\mathbf{R}^2} \langle v\rangle^{1-l} dv d\tau < \infty$$

because $l > 3$. It follows that I^+ is locally bounded and therefore so are $|\partial_x E|$ and $|\partial_x B|$.

Finally, we take derivatives of the Vlasov equation with respect to x, v_1, and v_2 to obtain

$$\begin{aligned}
L_\pm \partial_x f_\pm &= \mp(\partial_x E + \hat{v}\times\partial_x B)\cdot\nabla_v f_\pm,\\
L_\pm \partial_{v_1} f_\pm &= -\frac{1+v_2^2}{\langle v\rangle^3}\partial_x f_\pm \mp \frac{\partial(\hat{v})}{\partial v_1}\times B\cdot\nabla_v f,\\
L_\pm \partial_{v_2} f_\pm &= -\frac{v_1 v_2}{\langle v\rangle^3}\partial_x f_\pm \mp \frac{\partial(\hat{v})}{\partial v_2}\times B\cdot\nabla_v f,
\end{aligned}$$

where $L_\pm = \partial_t + \hat{v}_1\partial_x \pm (E+\hat{v}\times B)\cdot\nabla_v$. Let us also denote $\|\partial f\|_{1,A} = \sum_\pm \int_{|x|\le A}\int(|\partial_x f_\pm| + |\nabla_v f_\pm|)dvdx$. Then these equations directly lead to the local L^1 estimate

$$\begin{aligned}
&\|\partial f(t)\|_{1,A} \le \|\partial f(0)\|_{1,A+t}\\
&+C\int_0^t (1 + |\partial_x E(\tau)|_{\infty,A+t-\tau} + |\partial_x B(\tau)|_{\infty,A+t-\tau})\,\|\partial f(\tau)\|_{1,A+t-\tau} d\tau.
\end{aligned}$$

Since E and B are locally bounded in $W^{1,\infty}$, Gronwall's inequality implies the boundedness of $\|\partial f(t)\|_{1,A}$ for bounded t and A. Upon passage to the limit we obtain $f_\pm \in L^\infty(\Omega_t; BV(\Omega_x\times\mathbf{R}^2))$ for any Ω_t and Ω_x.

The uniqueness proof is standard. □

For $1\frac{1}{4}$RVM with periodic boundary conditions, we have the following theorem.

THEOREM 5. *Let $f_\pm^0(x,v_1,v_2) = f_\pm^0(x,v_1,-v_2)$. Let $f_\pm^0 \in BV(\mathbf{R}_P\times\mathbf{R}^2)$, $f_\pm^0 \ge 0$, $\langle v\rangle^l f_\pm^0 \in L^\infty(\mathbf{R}_P\times\mathbf{R}^2)$ for some $l > 2$, $E_1^0 \in W^{1,\infty}(\mathbf{R}_P)$,*

$$\int_0^P\int_{\mathbf{R}^2}(f_+^0 - f_-^0)dvdx = 0, \qquad \partial_x E_1^0 = \int_{\mathbf{R}^2}(f_+^0 - f_-^0)dv.$$

Then there exists a unique solution, of period P in x, with initial data $[f_+^0, f_-^0, E_1^0]$ such that $f_\pm \in L^\infty_{loc}(\mathbf{R}; BV(\mathbf{R}_P\times\mathbf{R}^2))$, $\langle v\rangle^l f_\pm \in L^\infty_{loc}(\mathbf{R}; L^\infty(\mathbf{R}_P\times\mathbf{R}^2))$, and $E_1 \in L^\infty_{loc}(\mathbf{R}; W^{1,\infty}(\mathbf{R}_P))$.

Proof. If $l > 3$, let $u = [f_+, f_-, E_1, E_2, B]$ be the solution of $1\frac{1}{2}$RVM with the initial data $[f_+^0, f_-^0, E_1^0, 0, 0]$. Let $\check{f}_\pm(t,x,v_1,v_2) = f_\pm(t,x,v_1,-v_2)$. It is easy to verify

that $[\check{f}_+, \check{f}_-, E_1, -E_2, -B]$ is another solution of $1\frac{1}{2}$RVM with the same initial data. By uniqueness in Theorem 4, they are equal. Therefore f_+ and f_- are even functions of v_2, and $E_2 \equiv B \equiv 0$. Thus Theorem 5 is a special case of Theorem 4. On the other hand, if $2 < l \leq 3$, we can prove our theorem directly just as in Theorem 4 except that all the discussion of E_2 and B can be eliminated. □

REFERENCES

[BGK] I. Bernstein, J. Greene, and M. Kruskal, *Exact nonlinear plasma oscillations*, Phys. Rev., 108 (1957), pp. 546–550.

[G1] Y. Guo, *Stable magnetic equilibria in collisionless plasmas*, Comm. Pure Appl. Math., L (1997), pp. 0891–0933.

[G2] Y. Guo, *Stable magnetic equilibria in a symmetric plasma*, Comm. Math. Phys., 200 (1999), pp. 211–247.

[GR] Y. Guo and C. G. Ragazzo, *On steady states in a collisionless plasma*, Comm. Pure Appl. Math., XLIX (1996), pp. 1145–1174.

[GS1] Y. Guo and W. Strauss, *Nonlinear instability of double-humped equilibria*, Ann. Inst. H. Poincaré Anal. Non. Linéaire, 12 (1995), pp. 339–352.

[GS2] Y. Guo and W. Strauss, *Instability of periodic BGK equilibria*, Comm. Pure Appl. Math., XLVIII (1995), pp. 861–894.

[GS3] Y. Guo and W. Strauss, *Relativistic unstable periodic BGK waves*, Comput. Appl. Math., 18 (1999), pp. 87–122.

[GS4] Y. Guo and W. Strauss, *Unstable BGK solitary waves and collisionless shocks*, Comm. Math. Phys., 195 (1998), pp. 267–293.

[GSc] R. Glassey and J. Schaeffer, *On the 'one and one-half-dimensional' relativistic system*, Math. Methods Appl. Sci., 13 (1990), pp. 169–179.

[P] O. Penrose, *Electrostatic instability of a non-Maxwellian plasma*, Phys. Fluids., 3 (1960), pp. 258–265.

[Sh] Y. Shizuta, *On the classical solutions of the Boltzmann equation*, Comm. Pure Appl. Math., 36 (1983), pp. 705–754.

[St] S. Steinberg, *Meromorphic families of compact operators*, Arch. Rational Mech. Anal., 31 (1968), pp. 372–379.

[V] I. Vidav, *Spectra of perturbed semigroups with applications to transport theory*, J. Math. Anal. Appl., 30 (1970), pp. 264–279.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1115–1135

ON THE ZERO RELAXATION LIMIT FOR A SYSTEM MODELING THE MOTIONS OF A VISCOELASTIC SOLID*

WEN SHEN†, ASLAK TVEITO†, AND RAGNAR WINTHER†

Abstract. We consider a simple model of the motions of a viscoelastic solid. The model consists of a two-by-two system of conservation laws including a strong relaxation term. We establish the existence of a BV-solution of this system for any positive value of the relaxation parameter. We also show that this solution is stable with respect to the perturbations of the initial data in L^1. By deriving the uniform bounds, with respect to the relaxation parameter, on the total variation of the solution, we obtain the convergence of the solutions of the relaxation system towards the solutions of a scalar conservation law as the relaxation parameter δ goes to zero. Due to the Lip^+ bound on the solutions of the relaxation system, an estimate on the rate of convergence towards equilibrium is derived. In particular, an $\mathcal{O}(\sqrt{\delta})$ bound on the L^1-error is established.

Key words. hyperbolic conservation laws, relaxation terms, nonequilibrium, convergence towards equilibrium, viscoelasticity, finite difference schemes

AMS subject classifications. 35L65, 65M99

PII. S003614109731984X

1. Introduction.

In this paper we study the following system of conservation laws:

$$u_t + \sigma_x = 0, \tag{1.1}$$
$$(\sigma - f(u))_t + \frac{1}{\delta}(\sigma - \mu f(u)) = 0,$$

where the parameters μ and δ satisfy $0 < \mu < 1$ and $0 < \delta \ll 1$. Here μ is a fixed parameter, while we are, in particular, interested in the limit as the relaxation parameter δ tends to zero.

If $\delta \to 0$, we formally obtain the equilibrium relation

$$\bar{\sigma} = \mu f(\bar{u}),$$

and hence the equilibrium model

$$\bar{u}_t + \mu f(\bar{u})_x = 0. \tag{1.2}$$

The purpose of this paper is to study the limit process rigorously. We will prove that under proper conditions on the initial data, the solutions of the nonequilibrium model converge to the solutions of the equilibrium model in L^1, uniformly in δ at a rate of $\mathcal{O}(\sqrt{\delta})$.

The system (1.1) arises in the modeling of motions of a viscoelastic solid, where the relaxation phenomenon presents the strength of memory. The Riemann problem for the system with $\delta = 1$ is studied by Greenberg and Hsiao [4]. The zero relaxation limit

*Received by the editors April 11, 1997; accepted for publication (in revised form) April 29, 1998; published electronically August 26, 1999. This research was supported by the Norwegian Research Council (NFR), program 110673/420, at the Department of Applied Mathematics, SINTEF, Oslo, Norway.

http://www.siam.org/journals/sima/30-5/31984.html

†Department of Informatics, P.O. Box 1080, Blindern, University of Oslo, N-0316 Oslo, Norway (wens@ifi.uio.no, aslak@ifi.uio.no, ragnar@ifi.uio.no).

of this viscoelasticity model with vanishing memory is analyzed in the fundamental paper of Chen and Liu [1], where nonlinear stability in the zero relaxation limit is established for the model. This is achieved by first deriving energy estimates from proper construction of entropy pairs, and then applying the theory of compensated compactness. More recent results can be found in the paper by Chen, Levermore, and Liu [2]. In this paper, we will establish similar results, but in the BV-framework. For any positive values of the relaxation parameter, we will prove the existence of a BV-solution of the system. The bound on the total variation of the solution, and a proper stability estimate with respect to perturbations of the initial data in L^1, are both independent of the relaxation parameter. Furthermore, a uniform Lip^+ bound, similar to Oleinik's entropy condition (cf. [12]), is obtained. By following the framework of Tadmor, Nessyahu, and Kurganov [15, 11, 6], this bound is used to establish an $\mathcal{O}(\sqrt{\delta})$ estimate for the L^1 difference between the solution of the relaxation system (1.1) and the solution of the equilibrium model (1.2).

Hyperbolic conservation laws with relaxation terms arise in modeling of many physical phenomena, such as chromatography, traffic modeling, water waves, and viscoelasticity (see, e.g., the book of Whitham [17]). General relaxation effect was analyzed by Liu [8], and the convergence was studied by Natalini [10]. For a system modeling chromatography, convergence and rate of convergence towards equilibrium are proved (cf., [13, 16] for the 1D case and [14] for the 2D case). Sharper estimates on the rate of convergence for this model have been recently derived by Kurganov and Tadmor [6]. The approach here resembles the techniques used in [6, 13, 16]. The same model problem is also studied independently by Yong [18] and Luo and Natalini [9]. However, these papers do not derive a rate for the convergence to equilibrium.

The structure of the paper is as follows. In section 2, we give the preliminaries for the model, and we also state the main results of the paper. Then the properties of the finite difference solutions are studied in section 3, where we establish the uniform bound, the TV bound, and the bound on the deviation from the equilibrium state. In section 4, we prove that the limit of the finite difference solution is the entropy solution of the system, and the stability in L^1 is then proved by Kruzkov-type arguments. Finally, the proof of the convergence of the solution of the nonequilibrium model towards the solution of the equilibrium model is given in section 5.

2. Preliminaries and statement of the main results. In this section, we will give the preliminaries of the paper and state the main result. Throughout this paper we will assume that the flux function $f = f(u)$ is a smooth function with the following properties:

$$f(0) = 0, \quad f'(u) > 0, \quad f''(u) \geq 0 \quad \text{for all } u \geq 0.$$

We introduce the variable $v = f(u) - \sigma$ such that $u = g(\sigma + v)$, where the function $g = f^{-1}$. Under the assumption that $u \geq 0$, we obtain a reformulation of the system (1.1):

$$\begin{aligned} g(\sigma + v)_t + \sigma_x &= 0, \\ v_t &= \frac{1}{\delta} R(\sigma, v), \end{aligned} \tag{2.1}$$

where $R(\sigma, v) = ((1 - \mu)\sigma - \mu v)$. The associated equilibrium model is

$$g\left(\frac{\bar{\sigma}}{\mu}\right)_t + \bar{\sigma}_x = 0. \tag{2.2}$$

We observe that the "reaction function" R has the monotonicity property

$$R(\sigma, v)(\operatorname{sgn}(\sigma) - \operatorname{sgn}(v)) \geq 0. \tag{2.3}$$

We seek solutions of (2.1) in the *state space*

$$\mathcal{S} = \{(\sigma, v) : 0 \leq \sigma \leq \mu \ , 0 \leq v \leq 1 - \mu\} \tag{2.4}$$

and solutions of (2.2) in $[0, \mu]$. For a scalar function $u(x)$, let $TV(u)$ denote the total variation defined as

$$TV(u) := \sup_{h \neq 0} \int_{\mathbb{R}} \frac{|u(x+h) - u(x)|}{h} \, dx,$$

and the L^1 norm is defined as

$$\|u\|_{L^1} := \int_{\mathbb{R}} |u(x)| \, dx.$$

Furthermore, we define

$$\operatorname{Lip}^+(u) := \max\left(0, \ \operatorname{ess} \sup_{x \neq y} \frac{u(x) - u(y)}{x - y}\right).$$

Let $p = R(\sigma, v)$ denote the residual. We assume the initial data (σ^0, v^0) satisfies the following requirements:

$$\begin{aligned}
&\text{i) } (\sigma^0(x), v^0(x)) \in \mathcal{S}, \quad \forall x \in \mathbb{R}, \\
&\text{ii) } TV(\sigma^0) + TV(v^0) \leq M, \\
&\text{iii) } \left\|p^0\right\|_{L^1} \leq M\delta, \\
&\text{iv) } \sigma^0(\pm\infty) = v^0(\pm\infty) = 0, \\
&\text{v) } \operatorname{Lip}^+(\sigma^0) \leq M, \ \ \operatorname{Lip}^+(v^0) \leq M.
\end{aligned} \tag{2.5}$$

Here, and throughout this paper, M denotes a generic positive finite constant independent of δ and the grid parameters $(\Delta x, \Delta t)$. Let $G = G(\sigma, v, k, q)$ be defined as

$$G(\sigma, v, k, q) = \frac{g(\sigma + v) - g(k + q)}{(\sigma + v) - (k + q)},$$

and for any $T > 0$, let $\mathcal{D}_+(T)$ be the set of all nonnegative $\mathcal{C}^\infty$-functions with compact support in $\mathbb{R} \times [0, T]$. Then the entropy solutions of (2.1) are defined as follows.

DEFINITION 2.1. *Let (σ^0, v^0) be the initial data of* (2.1) *which satisfies the assumptions in* (2.5). *Then a pair of functions (σ, v) is called the entropy solution of* (2.1) *with initial data (σ^0, v^0) if the following requirements are satisfied:*

i) $(\sigma, v) \in \mathcal{S}, \quad \forall (x, t) \in \mathbb{R} \times \mathbb{R}_0^+$,
ii) $TV(\sigma(\cdot, t)) + TV(v(\cdot, t)) \leq M, \quad \forall t \geq 0$,
iii) $\|\sigma(\cdot, t) - \sigma(\cdot, \tau)\|_{L^1} + \|v(\cdot, t) - v(\cdot, \tau)\|_{L^1} \leq M|t - \tau|, \quad \forall t, \tau \geq 0$,
iv) $Lip^+(\sigma(\cdot, t)) \leq M, \ Lip^+(v(\cdot, t)) \leq M, \quad \forall t \geq 0$,
v) *for any $(k, q) \in \mathcal{S}$ and any $\phi \in \mathcal{D}_+(T)$, the following inequality is valid for all $T > 0$:*

$$
\begin{aligned}
(2.6)\quad & \int_0^T \int_{\mathbb{R}} [G(\sigma,v,k,q)(|\sigma-k|+|v-q|)\phi_t + |\sigma-k|\phi_x]\,dx\,dt \\
& + \int_{\mathbb{R}} G(\sigma^0,v^0,k,q)(\left|\sigma^0-k\right|+\left|v^0-q\right|)\phi(x,0)\,dx \\
& - \int_{\mathbb{R}} G(\sigma(x,T),v(x,T),k,q)(|\sigma(x,T)-k|+|v(x,T)-q|)\ \phi(x,T)\,dx \\
& + M\int_0^T \int_{\mathbb{R}} [|v-q|-(v-q)sgn(\sigma-k)]\,\phi\,dx\,dt \\
& \geq \frac{1}{\delta}\int_0^T \int_{\mathbb{R}} G(\sigma,v,k,q)R(\sigma,v)[sgn(\sigma-k)-sgn(v-q)]\phi\,dx\,dt.
\end{aligned}
$$

Note that the entropy inequality in (2.6) is the weak formulation of an inequality of the form

$$\mathcal{E}_t + \mathcal{F}_x \leq -\frac{1}{\delta}\mathcal{G} + M\mathcal{H},$$

where

$$
\begin{aligned}
\mathcal{E} &= [G(\sigma,v,k,q)\big(|\sigma-k|+|v-q|\big)], \\
\mathcal{F} &= |\sigma-k|, \\
\mathcal{G} &= G(\sigma,v,k,q)R(\sigma,v)[\mathrm{sgn}(\sigma-k)-\mathrm{sgn}(v-q)], \\
\mathcal{H} &= |v-q|-(v-q)\mathrm{sgn}(\sigma-k).
\end{aligned}
$$

Remarks. In order to motivate the weak entropy formulation above, let us assume that (σ,v) and $(\bar\sigma,\bar v)$ are two smooth solutions of the system (2.1). The errors, $\sigma-\bar\sigma$ and $v-\bar v$, will then be governed by the system

$$
\begin{aligned}
[G\big((\sigma-\bar\sigma)+(v-\bar v)\big)]_t + (\sigma-\bar\sigma)_x &= 0, \\
(v-\bar v)_t &= \frac{1}{\delta}R,
\end{aligned}
$$

where $G = G(\sigma,v,\bar\sigma,\bar v)$ and $R = R(\sigma-\bar\sigma, v-\bar v)$. The system can also be rewritten as

$$
\begin{aligned}
G_t\,(\sigma-\bar\sigma) + G\,(\sigma-\bar\sigma)_t + (\sigma-\bar\sigma)_x &= -G_t\,(v-\bar v) - \frac{1}{\delta}GR, \\
G(v-\bar v)_t + G_t\,(v-\bar v) &= G_t\,(v-\bar v) + \frac{1}{\delta}GR.
\end{aligned}
$$

By multiplying the first equation above by $\mathrm{sgn}(\sigma-\bar\sigma)$ and the second one by $\mathrm{sgn}(v-\bar v)$, and summing, we obtain

$$
\begin{aligned}
(2.7)\quad & [G(|\sigma-\bar\sigma|+|v-\bar v|)]_t + (|\sigma-\bar\sigma|)_x \\
& = G_t[|v-\bar v|-(v-\bar v)\mathrm{sgn}(\sigma-\bar\sigma)] - \frac{1}{\delta}GR(\mathrm{sgn}(\sigma-\bar\sigma)-\mathrm{sgn}(v-\bar v)).
\end{aligned}
$$

If the function $G = G(x,t)$ satisfies a one-sided Lipschitz condition of the form

$$G_t(x,t) \leq M, \tag{2.8}$$

then clearly (2.7) implies that

$$[G(|\sigma-\bar\sigma|+|v-\bar v|)]_t+(|\sigma-\bar\sigma|)_x \le M\,[|v-\bar v|-(v-\bar v)\mathrm{sgn}(\sigma-\bar\sigma)]-\frac{1}{\delta}GR(\mathrm{sgn}\ (\sigma-\bar\sigma)-\mathrm{sgn}(v-\bar v)). \tag{2.9}$$

The weak entropy formulation above is motivated from this differential inequality. We also note that since $G\ge 0$, it follows from (2.3) and (2.9) that

$$[G(|\sigma-\bar\sigma|+|v-\bar v|)]_t+(|\sigma-\bar\sigma|)_x \le 2M|v-\bar v|.$$

This formal inequality indicates the continuous dependence result which will be established rigorously in this paper.

The motivation for the entropy formulation above relies on the one-sided bound (2.8). Since

$$G(\sigma,v,\bar\sigma,\bar v)=\int_0^1 g'(\theta(\sigma+v)+(1-\theta)(\bar\sigma+\bar v))\,d\theta,$$

and

$$(g'(\sigma+v))_t=-\frac{g''(\sigma+v)}{g'(\sigma+v)}\sigma_x\le M\sigma_x,$$

the bound (2.8) will follow from an estimate of the form

$$\mathrm{Lip}^+(\sigma(\cdot,t)),\mathrm{Lip}^+(\bar\sigma(\cdot,t))\le M.$$

As we shall see below, this property for solutions of the system (2.1) will essentially follow from the corresponding assumption (2.5v) on the initial data. This ends our discussion on the motivation for the weak entropy formulation.

For the scalar equilibrium equation, the entropy solutions are defined in the sense of Kruzkov [5]. For a given $T>0$, the entropy solutions satisfy the following inequality for any $k\in\mathcal{S}$ and any $\phi\in\mathcal{D}_+(T)$,

$$\int_0^T\int_{\mathbb{R}}\left(\left|g\left(\frac{\bar\sigma}{\mu}\right)-g\left(\frac{k}{\mu}\right)\right|\phi_t+|\sigma-k|\phi_x\right)\,dx\,dt + \int_{\mathbb{R}}\left[\left|g\left(\frac{\bar\sigma^0}{\mu}\right)-g\left(\frac{k}{\mu}\right)\right|\phi(x,0)-\left|g\left(\frac{\bar\sigma(x,T)}{\mu}\right)-g\left(\frac{k}{\mu}\right)\right|\phi(x,T)\right]\,dx\ge 0.$$

Our main tool in analyzing the system will be a finite difference scheme derived from the formulation (2.1). Let Δt and Δx denote the steplengths in the t and x directions, respectively. We consider a semi-implicit difference scheme of the form

$$\frac{g\left(\sigma_j^{n+1}+v_j^{n+1}\right)-g\left(\sigma_j^n+v_j^n\right)}{\Delta t}+\frac{\sigma_j^n-\sigma_{j-1}^n}{\Delta x}=0, \qquad \frac{v_j^{n+1}-v_j^n}{\Delta t}=\frac{1}{\delta}R\left(\sigma_j^{n+1},v_j^{n+1}\right). \tag{2.10}$$

Here σ_j^n and v_j^n denote approximations of $\sigma(x,t)$ and $v(x,t)$ over the gridblocks

$$B_j^n=[x_{j-1/2},x_{j+1/2})\times[t_n,t_{n+1}),$$

where $x_j = j\Delta x$ and $t_n = n\Delta t$. Let $u_m = g(1) > 0$, and let

$$M_f = \max_{u\in[0,u_m]} f'(u) = \left(\min_{\theta\in[0,1]} g'(\theta)\right)^{-1}.$$

Throughout the paper we shall assume that the CFL-condition

$$\lambda M_f \le 1 \tag{2.11}$$

is satisfied, where $\lambda \equiv \Delta t/\Delta x$ is the mesh ratio which we assume to be a constant. The discrete initial data is taken to be the cell averages

$$\sigma_i^0 := \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \sigma^0(x)\,dx, \qquad v_i^0 := \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} v^0(x)\,dx.$$

The total variation of a grid function u_i is defined as

$$TV(u) := \sum_i |u_i - u_{i-1}|,$$

and the discrete L^1-norm is

$$\|u\|_{L^1} := \Delta x \sum_i |u_i|.$$

We assume that the following requirements are satisfied:

$$\begin{aligned}
&\text{i) } (\sigma_i^0, v_i^0) \in \mathcal{S}, \quad \forall j,\\
&\text{ii) } TV(\sigma^0) + TV(v^0) \le M,\\
&\text{iii) } \left\|p^0\right\|_{L^1} \le M\delta,\\
&\text{iv) } \sigma_{\pm\infty}^0 = v_{\pm\infty}^0 = 0,\\
&\text{v) } \sup_j(\sigma_j^0 - \sigma_{j-1}^0) \le M\Delta t,\ \sup_j(v_j^0 - v_{j-1}^0) \le M\Delta t, \quad \forall j.
\end{aligned} \tag{2.12}$$

Note that the requirement (v) follows directly from the assumption in (2.5v).

The existence of an entropy solution of the Cauchy problem can be obtained based on the properties of the finite different solutions of the scheme (2.10). Furthermore, the well-posedness of the initial value problem, independent of δ, is also proved.

THEOREM 2.2. *Let (σ^0, v^0) be the initial data of* (2.1) *satisfying the conditions* (2.5)*, and let (σ_i^0, v_i^0) be the discrete initial data for scheme* (2.10)*. Let $(\sigma_\Delta, v_\Delta)$ be the piecewise constant representation of the grid data (σ_i^n, v_i^n) generated by scheme* (2.10)*. Then the family $\{(\sigma_\Delta, v_\Delta)\}$ of approximate solutions converge in $(L^1_{loc}(\mathbb{R}\times\mathbb{R}_0^+))^2$ towards a pair of functions (σ, v) as the grid parameters $(\Delta x, \Delta t)$ tend to zero. The limit is the unique entropy solution which satisfies the requirements in Definition* 2.1*, and the following bounds are valid:*

$$\begin{aligned}
&\|p(\cdot,t)\|_{L^1} \le M\delta,\\
&Lip^+(\sigma(\cdot,t)) \le M, \quad Lip^+(v(\cdot,t)) \le M.
\end{aligned}$$

Moreover, the solution is stable with respect to perturbations in initial data in the following sense: Let $(\bar\sigma, \bar v)$ be another entropy solution of (2.1) *with initial data $(\bar\sigma^0, \bar v^0)$. Then the following bound holds for all $t > 0$:*

$$\|\sigma(\cdot,t) - \bar\sigma(\cdot,t)\|_{L^1} + \|v(\cdot,t) - \bar v(\cdot,t)\|_{L^1} \le \bar M e^{Mt}\left[\left\|\sigma^0 - \bar\sigma^0\right\|_{L^1} + \left\|v^0 - \bar v^0\right\|_{L^1}\right],$$

where $\bar{M}$ and M are finite constants independent of δ.

This theorem eventually leads to the main result of this paper, i.e., the convergence of the solutions of the nonequilibrium system towards the solutions of the equilibrium equation as δ tends to zero, and an estimate of the rate of convergence. The error estimates are derived by following the framework of Tadmor, Nessyahu, and Kurganov [15, 11, 6]. Hence, we estimate the Lip$'$-norm of the error. For any function $\phi \in L^1$ with $\int \phi = 0$, we define

$$\|\phi\|_{\mathrm{Lip}'} := \sup_{\psi} \frac{\int_{\mathbb{R}} \phi\psi dx}{\|\psi\|_{W^{1,\infty}}}.$$

Here the supremum is taken over all smooth functions ψ with compact support and

$$\|\psi\|_{W^{1,\infty}} := \max\left(\|\psi\|_{L^\infty}, ||\psi||_{Lip}\right).$$

The following convergence result will be proved in section 5.

THEOREM 2.3. *Let (σ^0, v^0) and $\bar{\sigma}^0$ be the initial data for* (2.1) *and* (2.2)*, respectively. We assume that the initial data (σ^0, v^0) for the nonequilibrium system satisfies the requirements in* (2.5) *and that $\bar{\sigma}^0 = \sigma^0$. Let $(\sigma_\delta, v_\delta)$ be the entropy solution of* (2.1) *with initial data (σ^0, v^0) and $\bar{\sigma}$ the corresponding entropy solution of* (2.2)*. For each $T > 0$ there is a constant M, independent of δ, such that*

$$\|u_\delta(\cdot, t) - \bar{u}(\cdot, t)\|_{Lip'} \le M\delta, \quad 0 \le t \le T,$$

where $u_\delta = g(\sigma_\delta + v_\delta)$ and $\bar{u} = g(\frac{\bar{\sigma}}{\mu})$.

We note that the variables $(u_\delta, \sigma_\delta)$ and $(\bar{u}, \bar{\sigma})$ in the theorem above correspond to the solutions of the original models (1.1) and (1.2). The following corollary is a consequence of Theorem 2.3.

COROLLARY 2.4. *Let $(u_\delta, \sigma_\delta)$ and $(\bar{u}, \bar{\sigma})$ be as stated in Theorem* 2.3*. For each $T > 0$ there is a constant M, independent of δ, such that for any $p \in [1, \infty)$*

$$\|u_\delta(\cdot, t) - \bar{u}(\cdot, t)\|_{L^p} \le M\delta^{\frac{1}{2p}}, \quad 0 \le t \le T.$$

Furthermore,

$$\|\sigma_\delta(\cdot, t) - \bar{\sigma}(\cdot, t)\|_{L^1} \le M\sqrt{\delta}, \quad 0 \le t \le T.$$

3. Existence of a weak solution. The purpose of this section is to use the finite difference scheme (2.10) to establish the existence of weak solutions of Cauchy problem for (2.1) (or (1.1)). We first show that the finite difference solution is well defined.

LEMMA 3.1. *Assume that $\{\sigma_j^0\}$ and $\{v_j^0\}$ for $j \in \mathcal{Z}$ are given. Then the solutions $\{\sigma_j^n\}$ and $\{v_j^n\}$ are uniquely determined by* (2.10) *for all $j \in \mathcal{Z}$ and $n \ge 0$.*

Proof. Assume that $\{\sigma_j^n\}$ and $\{v_j^n\}$ are computed. Let

$$r_j^n = g\left(\sigma_j^n + v_j^n\right) - \lambda\left(\sigma_j^n - \sigma_{j-1}^n\right).$$

The solutions $\{\sigma_j^{n+1}\}$ and $\{v_j^{n+1}\}$ then satisfy the linear system

$$A\begin{pmatrix} \sigma_j^{n+1} \\ v_j^{n+1} \end{pmatrix} = \begin{pmatrix} f\left(r_j^n\right) \\ v_j^n \end{pmatrix},$$

where the 2×2 matrix A is given by

$$A = \begin{pmatrix} 1 & 1 \\ -(1-\mu)\frac{\Delta t}{\delta} & 1+\mu\frac{\Delta t}{\delta} \end{pmatrix}.$$

Since $\det(\mathrm{A}) = 1 + \frac{\Delta t}{\delta} > 0$, the results follows by induction. ∎

The following results show that the state space $\mathcal{S}$ defined in (2.4), is an invariant region for the scheme (2.10).

LEMMA 3.2. *Assume $(\sigma_j^0, v_j^0) \in \mathcal{S}$ for all $j \in \mathcal{Z}$. Then $(\sigma_j^n, v_j^n) \in \mathcal{S}$ for all $j \in \mathcal{Z}$ and $n \geq 0$.*

Proof. For given $\bar{\sigma}, \sigma_L$ and $\bar{v}$, let (σ, v) be the unique solution of the system

$$\begin{aligned} g(\sigma + v) &= g(\bar{\sigma} + \bar{v}) - \lambda\left(\bar{\sigma} - \sigma_L\right), \\ \left(1 + \frac{\Delta t}{\delta}\mu\right) v - \frac{\Delta t}{\delta}(1-\mu)\sigma &= \bar{v}. \end{aligned} \tag{3.1}$$

This system defines functions $\sigma = \sigma(\bar{\sigma}, \sigma_L, \bar{v})$ and $v = v(\bar{\sigma}, \sigma_L, \bar{v})$. Furthermore, $\sigma_j^{n+1} = \sigma(\sigma_j^n, \sigma_{j-1}^n, v_j^n)$ and $v_j^{n+1} = v(\sigma_j^n, \sigma_{j-1}^n, v_j^n)$. Hence, the lemma can be established by studying the functions σ and v.

Assume that $(\bar{\sigma}, \bar{v}) \in \mathcal{S}$ and $\sigma_L \in [0, \mu]$. By differentiating the system (3.1) with respect to $\bar{\sigma}$ and by using the CFL-condition (2.11), we obtain

$$\begin{aligned} g'(\sigma + v)\left(\frac{\partial \sigma}{\partial \bar{\sigma}} + \frac{\partial v}{\partial \bar{\sigma}}\right) &= g'(\bar{\sigma} + \bar{v}) - \lambda > 0, \\ \left(1 + \frac{\Delta t \mu}{\delta}\right)\frac{\partial v}{\partial \bar{\sigma}} &= \frac{\Delta t}{\delta}(1-\mu)\frac{\partial \sigma}{\partial \bar{\sigma}}. \end{aligned}$$

From this we easily conclude that $\frac{\partial \sigma}{\partial \bar{\sigma}}, \frac{\partial v}{\partial \bar{\sigma}} > 0$, and by a similar calculation we also obtain $\frac{\partial \sigma}{\partial \sigma_L}, \frac{\partial v}{\partial \sigma_L} > 0$.

Assume now that $\sigma_L = \bar{\sigma}$. Then we obtain from (3.1) that

$$\sigma + v = \bar{\sigma} + \bar{v},$$

and hence

$$\frac{\partial \sigma}{\partial \bar{v}} + \frac{\partial v}{\partial \bar{v}} = 1.$$

Furthermore, from the second equation of (3.1) we have

$$\frac{\Delta t}{\delta}\mu\frac{\partial v}{\partial \bar{v}} = \left(1 + \frac{\Delta t}{\delta}(1-\mu)\right)\frac{\partial \sigma}{\partial \bar{v}},$$

and hence we can conclude that

$$\frac{\partial \sigma}{\partial \bar{v}}(\bar{\sigma}, \bar{\sigma}, \bar{v}) > 0, \quad \frac{\partial v}{\partial \bar{v}}(\bar{\sigma}, \bar{\sigma}, \bar{v}) > 0.$$

From the monotonicity properties derived above we now have for $(\bar{\sigma}, \bar{v}) \in \mathcal{S}$ and $\sigma_L \in [0, \mu]$

$$\sigma(\bar{\sigma}, \sigma_L, \bar{v}) \geq \sigma(0, 0, \bar{v}) \geq \sigma(0, 0, 0) = 0$$

and

$$\sigma(\bar{\sigma}, \sigma_L, \bar{v}) \leq \sigma(\mu, \mu, \bar{v}) \leq \sigma(\mu, \mu, 1-\mu) = \mu.$$

Similarly, we obtain

$$0 \leq v(\bar{\sigma}, \sigma_L, \bar{v}) \leq 1-\mu,$$

and the invariance of $\mathcal{S}$ follows by induction. □

We let p_j^n denote the residual, i.e., $p_j^n = (1-\mu)\sigma_j^n - \mu v_j^n$.

LEMMA 3.3. *Assume that* $\|p^0\|_1$, $TV(\sigma^0)$ *and* $TV(v^0)$ *are finite. Then*

$$TV(\sigma^n) + TV(v^n) \leq TV(\sigma^0) + TV(v^0) \ . \tag{3.2}$$

Furthermore, there is a constant M_1*, depending only on* μ*,* g*,* $TV(\sigma^0)$*, and* $TV(v^0)$ *such that*

$$\frac{\|p^n\|_1}{\delta} \leq \max\left(M_1, \frac{\|p^0\|_1}{\delta}\right).$$

Proof. We first establish the total variation estimate. Let

$$a_j^n = \frac{\left(\sigma_j^{n+1} + v_j^{n+1}\right) - \left(\sigma_j^n + v_j^n\right)}{g\left(\sigma_j^{n+1} + v_j^{n+1}\right) - g\left(\sigma_j^n + v_j^n\right)}.$$

It follows from the monotonicity of g and the CFL-condition (2.11) that

$$0 \leq \lambda a_j^n \leq 1.$$

Furthermore, the difference scheme (2.10) can be written in the form

$$\begin{aligned} \sigma_j^{n+1} &= \sigma_j^n - \lambda a_j^n \left(\sigma_j^n - \sigma_{j-1}^n\right) - \frac{\Delta t}{\delta} R\left(\sigma_j^{n+1}, v_j^{n+1}\right), \\ v_j^{n+1} &= v_j^n + \frac{\Delta t}{\delta} R\left(\sigma_j^{n+1}, v_j^{n+1}\right). \end{aligned} \tag{3.3}$$

Hence, if we let

$$\alpha_j^n = \sigma_{j+1}^n - \sigma_j^n, \quad \beta_j^n = v_{j+1}^n - v_j^n,$$

we obtain

$$\begin{aligned} \alpha_j^{n+1} &= \alpha_j^n - \lambda a_{j+1}^n \alpha_j^n + \lambda a_j^n \alpha_{j-1}^n - \frac{\Delta t}{\delta} R\left(\alpha_j^{n+1}, \beta_j^{n+1}\right), \\ \beta_j^{n+1} &= \beta_j^n + \frac{\Delta t}{\delta} R\left(\alpha_j^{n+1}, \beta_j^{n+1}\right). \end{aligned} \tag{3.4}$$

By multiplying the first equation in (3.4) by sgn (α_j^{n+1}), the second equation by sgn (β_j^{n+1}), using the monotonicity property (2.3), and by summation with respect to j, we obtain

$$\sum_j \left(\left|\alpha_j^{n+1}\right| + \left|\beta_j^{n+1}\right|\right) \leq \sum_j \left(\left|\alpha_j^n\right| + \left|\beta_j^n\right|\right),$$

and this is exactly the total variation bound.

From (3.3) it also follows that

$$p_j^{n+1} = p_j^n - (1-\mu)\lambda a_j^n \left(\sigma_j^n - \sigma_{j-1}^n\right) - \frac{\Delta t}{\delta} p_j^{n+1}.$$

Therefore, it follows from the total variation estimate above that

$$\|p^{n+1}\|_1 \leq \|p^n\|_1 + M_1 \Delta t - \frac{\Delta t}{\delta}\|p^{n+1}\|_1,$$

and this implies that

$$\frac{\|p^{n+1}\|_1}{\delta} \leq \max\left(M_1, \frac{\|p^n\|_1}{\delta}\right).$$

This completes the proof of Lemma 3.3. ∎

We recall that the initial data satisfies

$$\|p^0\|_1 \leq M\delta,$$

where M is independent of δ and the grid parameters Δt and Δx. Hence, by induction, we have

$$\|p^n\|_1 \leq M\delta \quad \text{for all} \quad n \geq 0. \tag{3.5}$$

From the total variation estimate (3.2) and (3.5), we now obtain

$$\|\sigma^{n+1} - \sigma^n\|_1 + \|v^{n+1} - v^n\|_1 \leq M\Delta t,$$

and hence we obtain L^1-Lipschitz continuity with respect to time, i.e.,

$$\|\sigma^n - \sigma^m\|_1 + \|v^n - v^m\|_1 \leq M|n-m|\Delta t,$$

where M is independent of δ and the grid parameters.

4. Entropy solutions and stability in L^1. The purpose of this section is to derive bounds for $\mathrm{Lip}^+(\sigma)$ and $\mathrm{Lip}^+(v)$, which can be used to obtain stability results with respect to perturbations of the initial data which are independent of the relaxation parameter δ. The extra regularity results will technically be derived for the finite difference solutions (σ_j^n, v_j^n).

Define coefficients b_j^n by

$$b_j^n = \frac{a_{j+1}^n - a_j^n}{\alpha_j^{n+1} + \beta_j^{n+1} + \alpha_j^n + \beta_j^n},$$

where as above $\alpha_j^n = \sigma_{j+1}^n - \sigma_j^n$ and $\beta_j^n = v_{j+1}^n - v_j^n$. Observe that if we let $u_j^n = g(\sigma_j^n + v_j^n)$, then

$$a_j^n = \int_0^1 f'\left(u_j^n + \theta\left(u_j^{n+1} - u_j^n\right)\right)\, d\theta.$$

Hence, it follows from the monotonicity of f' and f that there is a positive constant M_b such that

$$0 < b_j^n \leq M_b.$$

We claim that for sufficiently small Δt and δ, the initial data (σ_j^0, v_j^0) of (2.10) satisfies the following one-side bound:

$$\sup_j \left\{(1-\mu)\alpha_j^0\ , \mu\beta_j^0\right\} \le (1-\mu)\mu^2 \frac{\Delta t}{2\delta + \mu\Delta t}. \tag{4.1}$$

Indeed, since $\alpha_j^0 \le \mu$ and $\beta_j^0 \le 1-\mu$ for all j, then by (2.12v), there exists a finite constant M^* and a sufficiently small Δt^* satisfying the relation $M^* \cdot \Delta t^* \le 1$ such that

$$\sup_j \left\{\alpha_j^0\right\} \le M^*\Delta t\mu, \quad \sup_j \left\{\beta_j^0\right\} \le M^*\Delta t(1-\mu),$$

for all $\Delta t \le \Delta t^*$. Then it follows that

$$\sup_j \left\{(1-\mu)\alpha_j^0, \mu\beta_j^0\right\} \le (1-\mu)\mu M^*\Delta t,$$

for all $\Delta t \le \Delta t^*$. By choosing δ sufficiently small, i.e.,

$$\delta \le \frac{\mu(1-M^*\Delta t)}{2M^*},$$

the relation (4.1) follows.

In order to derive the proper results for the solution of the finite difference scheme, we will need a strengthened CFL-condition. We will assume throughout this section that

$$\lambda\left(M_f + (2+\mu)M_b\right) \le 1. \tag{4.2}$$

LEMMA 4.1. *Assume that the initial data* (σ_j^0, v_j^0) *of* (2.10) *satisfies* (4.1) *for sufficiently small* δ *and* Δt. *Then*

$$\sup_j \left\{(1-\mu)\alpha_j^n, \mu\beta_j^n, 0\right\} \le \sup_j \left\{(1-\mu)\alpha_j^0, \mu\beta_j^0, 0\right\}.$$

Proof. Define function $\alpha = \alpha(\bar\alpha, \bar\beta, \alpha_L)$ and $\beta = \beta(\bar\alpha, \bar\beta, \alpha_L)$ implicitly by

$$\begin{aligned} \alpha &= \bar\alpha - \lambda a\left(\bar\alpha - \alpha_L\right) - \lambda b\left(\alpha + \beta + \bar\alpha + \bar\beta\right)\bar\alpha - \frac{\Delta t}{\delta}((1-\mu)\alpha - \mu\beta), \\ \beta &= \bar\beta + \frac{\Delta t}{\delta}((1-\mu)\alpha - \mu\beta). \end{aligned} \tag{4.3}$$

Here a and b are positive constants, bounded by M_f and M_b, respectively.

Recall that it follows from (3.4) that if $a = a_j^n$ and $b = b_j^n$, then $\alpha_j^{n+1} = \alpha(\alpha_j^n, \beta_j^n, \alpha_{j-1}^n)$ and $\beta_j^{n+1} = \beta(\alpha_j^n, \beta_j^n\ , \alpha_{j-1}^n)$. Recall also that Lemma 3.2 implies that $|\alpha_j^n| \le \mu$ and $|\beta_j^n| \le 1-\mu$.

We will first show that, under the assumptions that $|\bar\alpha|, |\alpha_L| \le \mu$, $|\bar\beta| \le 1-\mu$ and

$$\bar\alpha \le \mu^2 \frac{\Delta t}{2\delta + \mu\Delta t}, \tag{4.4}$$

the functions α and β are monotonically increasing in all three arguments. Observe that the second equation of (4.3) implies that

$$\beta = \frac{\delta}{\delta + \mu\Delta t}\bar\beta + \frac{(1-\mu)\Delta t}{\delta + \mu\Delta t}\alpha. \tag{4.5}$$

Hence we can eliminate β from the first equation. We obtain the equation

$$c\alpha = r, \tag{4.6}$$

where

$$c = c(\bar{\alpha}) = 1 + \lambda b \frac{\delta + \Delta t}{\delta + \mu \Delta t}\bar{\alpha} + \frac{(1-\mu)\Delta t}{\delta + \mu \Delta t} = (1 + \lambda b \bar{\alpha}) \frac{\delta + \Delta t}{\delta + \Delta t \mu}$$

and

$$r = r(\bar{\alpha}\ , \bar{\beta}\ , \alpha_L) = (1 - \lambda a)\bar{\alpha} + \lambda a \alpha_L - \lambda b \bar{\alpha}^2 - \lambda b \frac{2\delta + \mu \Delta t}{\delta + \mu \Delta t}\bar{\alpha}\bar{\beta} + \frac{\mu \Delta t}{\delta + \mu \Delta t}\bar{\beta}.$$

Note that since $\bar{\alpha} \geq -\mu$, it follows that

$$c \geq c(-\mu) \geq \frac{\delta + \Delta t}{\delta + \mu \Delta t}\left(1 - \mu \lambda M_b\right),$$

and hence (4.2) implies that $c > 0$. Observe that

$$\frac{\partial r}{\partial \alpha_L} = \lambda a > 0,$$

which implies that $\frac{\partial \alpha}{\partial \alpha_L} > 0$.

Similarly, by (4.2) and (4.4), we get

$$\frac{\partial r}{\partial \bar{\beta}} = \frac{\mu \Delta t - \lambda b(2\delta + \mu \Delta t)\bar{\alpha}}{\delta + \mu \Delta t} \geq \frac{\mu \Delta t}{\delta + \mu \Delta t}(1 - \lambda b \mu) \geq 0,$$

which implies that

$$\frac{\partial \alpha}{\partial \bar{\beta}} \geq 0.$$

Finally, we observe that

$$\begin{aligned} c\frac{\partial \alpha}{\partial \bar{\alpha}} &= \frac{\partial r}{\partial \bar{\alpha}} - \alpha \frac{dc}{d\bar{\alpha}} \\ &= (1 - \lambda a) - 2\lambda b \bar{\alpha} - \lambda b \frac{2\delta + \mu \Delta t}{\delta + \mu \Delta t}\bar{\beta} - \lambda b \frac{\delta + \Delta t}{\delta + \mu \Delta t}\alpha \\ &\geq (1 - \lambda a) - 2\lambda b \mu - \lambda b \frac{2\delta + \mu \Delta t}{\delta + \mu \Delta t}(1 - \mu) - \lambda b \frac{\delta + \Delta t}{\delta + \mu \Delta t}\mu. \end{aligned}$$

This implies that

$$c\frac{\partial \alpha}{\partial \bar{\alpha}} \geq 1 - \lambda(a + b(2 + \mu)).$$

Hence, it follows from (4.2) that

$$\frac{\partial \alpha}{\partial \bar{\alpha}} \geq 0.$$

We have therefore established that the function α is an increasing function in all three of its arguments. Furthermore, from (4.5) we easily derive that β has the corresponding property. We now use induction to complete the proof. Assume that

$$z^n \equiv \sup_j \left\{(1 - \mu)\alpha_j^n\ , \mu \beta_j^n\ , 0\right\} \leq z^0.$$

In particular, this implies that (cf. (4.4))

$$\alpha_j^n \le \mu^2 \frac{\Delta t}{2\delta + \mu\Delta t}.$$

Hence, the monotonicity property of α implies that

$$\alpha_j^{n+1} \le \alpha\left(\frac{z^n}{1-\mu}, \frac{z^n}{\mu}, \frac{z^n}{1-\mu}\right).$$

Furthermore, since $z^n \ge 0$,

$$c\left(\frac{z^n}{1-\mu}\right) \ge \frac{\delta + \Delta t}{\delta + \mu\Delta t}$$

and

$$r\left(\frac{z^n}{1-\mu}, \frac{z^n}{\mu}, \frac{z^n}{1-\mu}\right) \le \frac{z^n}{1-\mu} + \frac{\Delta t\, z^n}{\delta + \mu\Delta t} = \frac{z^n}{1-\mu}\left(\frac{\delta + \Delta t}{\delta + \mu\Delta t}\right).$$

We therefore obtain from (4.6) that

$$\alpha_j^{n+1} = \frac{r\left(\frac{z^n}{1-\mu}, \frac{z^n}{\mu}, \frac{z^n}{1-\mu}\right)}{c\left(\frac{z^n}{1-\mu}\right)} \le \frac{z^n}{1-\mu}.$$

Finally, from (4.5), we derive

$$\beta_j^{n+1} \le \beta\left(\frac{z^n}{1-\mu}, \frac{z^n}{\mu}, \frac{z^n}{1-\mu}\right) \le \frac{\delta}{\delta + \mu\Delta t}\frac{z^n}{\mu} + \frac{(1-\mu)\Delta t}{\delta + \mu\Delta t}\frac{z^n}{1-\mu} = \frac{z^n}{\mu}.$$

Hence, we conclude that $z^{n+1} \le z^n$. □

Next we will show that the finite difference solution satisfies a "discrete entropy inequality." Recall that the initial data (σ^0, v^0) satisfies a one-sided bound of the form (cf. (2.12v))

$$\sup_j \left\{\sigma_j^0 - \sigma_{j-1}^0, v_j^0 - v_{j-1}^0\right\} \le M\Delta t, \tag{4.7}$$

where $M > 0$ is a finite constant independent of δ and the mesh parameters. For $(\sigma, v), (k, q) \in \mathcal{S}$, we define

$$G(\sigma, v, k, q) = \frac{g(\sigma + v) - g(k + q)}{(\sigma + v) - (k + q)}.$$

Hence,

$$G(\sigma, v, k, q) \ge M_f^{-1} > 0.$$

For a fixed $(k, q) \in \mathcal{S}$, let

$$G_j^n = G\left(\sigma_j^n, v_j^n, k, q\right),$$

where $\{(\sigma_j^n, v_j^n)\}$ denotes the solution of the difference scheme (2.10). Observe that it follows from (2.10) that

$$G_j^{n+1} - G_j^n = -\lambda \, \frac{G_j^{n+1} - G_j^n}{\left(\sigma_j^{n+1} + v_j^{n+1}\right) - \left(\sigma_j^n + v_j^n\right)} \cdot \frac{f\left(u_j^{n+1}\right) - f\left(u_j^n\right)}{u_j^{n+1} - u_j^n} \left(\sigma_j^n - \sigma_{j-1}^n\right).$$

Therefore, since f is increasing and g is concave (because $g'' = -f''/(f')^3 \le 0$), it follows that there is a positive constant M, depending only on f (or g), such that

$$G_j^{n+1} - G_j^n \le M \max\left(0, \sigma_j^n - \sigma_{j-1}^n\right). \tag{4.8}$$

Hence, we obtain from (4.8), (4.7), and Lemma 4.1 that

$$G_j^{n+1} - G_j^n \le M \Delta t, \tag{4.9}$$

where $M > 0$ is independent of δ and the mesh parameters.

LEMMA 4.2. *There is a positive constant M, independent of δ and the mesh parameters such that for any $(k, q) \in \mathcal{S}$ the solution of* (2.10) *satisfies*

$$\begin{aligned} &G_j^{n+1}\left(\left|\sigma_j^{n+1} - k\right| + \left|v_j^{n+1} - q\right|\right) \\ &\quad \le G_j^n\left(\left|\sigma_j^n - k\right| + \left|v_j^n - q\right|\right) - \lambda\left(\left|\sigma_j^n - k\right| - \left|\sigma_{j-1}^n - k\right|\right) \\ &\quad -\frac{\Delta t}{\delta} G_j^n R\left(\sigma_j^{n+1}, v_j^{n+1}\right)\left[sgn\left(\sigma_j^{n+1} - k\right) - sgn\left(v_j^{n+1} - q\right)\right] \\ &\quad +M\Delta t\left[\left|v_j^{n+1} - q\right| - \left(v_j^{n+1} - q\right)\, sgn\, \left(\sigma_j^{n+1} - q\right)\right], \end{aligned}$$

where, as above, $G_j^n = G(\sigma_j^n, v_j^n, k, q)$.

Proof. Let $(k, q) \in \mathcal{S}$. From the first equation in (2.10) we directly obtain

$$\begin{aligned} G_j^{n+1}\left(\sigma_j^{n+1} - k\right) = G_j^n\left(\sigma_j^n - k\right) - \lambda\left(\sigma_j^n - \sigma_{j-1}^n\right) \\ - \left(G_j^{n+1} - G_j^n\right)\left(v_j^{n+1} - q\right) - G_j^n\left(v_j^{n+1} - v_j^n\right). \end{aligned}$$

Hence, by using the second equation of (2.10), this can be written in the form

$$\begin{aligned} G_j^{n+1}\left(\sigma_j^{n+1} - k\right) = G_j^n\left(\sigma_j^n - k\right) - \lambda\left[\left(\sigma_j^n - k\right) - \left(\sigma_{j-1}^n - k\right)\right] \\ - \left(G_j^{n+1} - G_j^n\right)\left(v_j^{n+1} - q\right) - \frac{\Delta t}{\delta} G_j^n R_j^{n+1}, \end{aligned} \tag{4.10}$$

where $R_j^{n+1} = R(\sigma_j^{n+1}, v_j^{n+1})$.

The next step in the derivation is to multiply (4.10) by sgn $(\sigma_j^{n+1} - k)$. Observe that since $0 < \lambda \le M_f^{-1} \le G_j^n$, the inequality

$$\begin{aligned} \left\{G_j^n\left(\sigma_j^n - k\right) - \lambda\left[\left(\sigma_j^n - k\right) - \left(\sigma_{j-1}^n - k\right)\right]\right\} \operatorname{sgn}\left(\sigma_j^{n+1} - k\right) \\ \le G_j^n\left|\sigma_j^n - k\right| - \lambda\left(\left|\sigma_j^n - k\right| - \left|\sigma_{j-1}^n - k\right|\right) \end{aligned}$$

holds. Hence, from (4.10), we obtain

$$\begin{aligned} G_j^{n+1}\left|\sigma_j^{n+1} - k\right| \le G_j^n\left|\sigma_j^n - k\right| - \lambda\left(\left|\sigma_j^n - k\right| - \left|\sigma_{j-1}^n - k\right|\right) \\ - \left[\left(G_j^{n+1} - G_j^n\right)\left(v_j^{n+1} - q\right) + \frac{\Delta t}{\delta} G_j^n R_j^{n+1}\right] \operatorname{sgn}\left(\sigma_j^{n+1} - k\right). \end{aligned} \tag{4.11}$$

Next, write the second equation of (2.10) in the form

$$G_j^{n+1}\left(v_j^{n+1}-q\right)=G_j^n\left(v_j^n-q\right)+\left(G_j^{n+1}-G_j^n\right)\left(v_j^{n+1}-q\right)+\frac{\Delta t}{\delta}G_j^nR_j^{n+1}.$$

Hence, if we multiply this equation by $\operatorname{sgn}\left(v_j^{n+1}-q\right)$ and add the result to (4.11) we obtain the inequality

$$\begin{aligned}G_j^{n+1}\left(\left|\sigma_j^{n+1}-k\right|+\left|v_j^{n+1}-q\right|\right)&\le G_j^n\left(\left|\sigma_j^n-k\right|+\left|v_j^n-q\right|\right)\\&-\lambda\left(\left|\sigma_j^n-k\right|-\left|\sigma_{j-1}^n-k\right|\right)\\&-\frac{\Delta t}{\delta}G_j^nR_j^{n+1}\left[\operatorname{sgn}\left(\sigma_j^{n+1}-k\right)-\operatorname{sgn}\left(v_j^{n+1}-q\right)\right]\\&+\left(G_j^{n+1}-G_j^n\right)\left(v_j^{n+1}-q\right)\left[\operatorname{sgn}\left(v_j^{n+1}-q\right)-\operatorname{sgn}\left(\sigma_j^{n+1}-k\right)\right].\end{aligned}\tag{4.12}$$

However, note that

$$0\le\left(v_j^{n+1}-q\right)\left[\operatorname{sgn}\left(v_j^{n+1}-q\right)-\operatorname{sgn}\left(\sigma_j^{n+1}-k\right)\right];$$

therefore, it follows from the one-sided bound (4.9) that

$$\begin{aligned}&\left(G_j^{n+1}-G_j^n\right)\left(v_j^{n+1}-q\right)\left[\operatorname{sgn}\left(v_j^{n+1}-q\right)-\operatorname{sgn}\left(\sigma_j^{n+1}-k\right)\right]\\&\quad\le M\Delta t\left[\left|v_j^{n+1}-q\right|-\left(v_j^{n+1}-q\right)\operatorname{sgn}\left(\sigma_j^{n+1}-q\right)\right],\end{aligned}$$

and hence the desired inequality follows from (4.12). □

Consider a real valued function $\mathcal{E}:\mathcal{S}\mapsto\mathbb{R}$ of the form

$$\mathcal{E}(\sigma,v)=\mathcal{L}(g(\sigma+v))+\int_{\mathcal{S}}P(k,q)G(\sigma,v,k,q)(|\sigma-k|+|v-q|)\,dk\,dq.$$

Here, $\mathcal{L}$ is a linear function and $P:\mathcal{S}\mapsto\mathbb{R}$ is a smooth, nonnegative function. Define, correspondingly,

$$\begin{aligned}\mathcal{F}(\sigma)&=\mathcal{L}(\sigma)+\int_{\mathcal{S}}P(k,q)|\sigma-k|\,dk\,dq,\\\mathcal{G}(\bar\sigma,\bar v,\sigma,v)&=\int_{\mathcal{S}}P(k,q)G(\bar\sigma,\bar v,k,q)R(\sigma,v)[\operatorname{sgn}(\sigma-k)-\operatorname{sgn}(v-q)]\,dk\,dq,\\\mathcal{H}(v)&=\int_{\mathcal{S}}P(k,q)\left[|v-q|-(v-q)\operatorname{sgn}(\sigma-k)\right]\,dk\,dq.\end{aligned}$$

It follows from (2.10) and by integrating the inequality of Lemma 4.2 that the solution of (2.10) satisfies the discrete entropy inequality

$$\begin{aligned}\mathcal{E}\left(\sigma_j^{n+1},v_j^{n+1}\right)&\le\mathcal{E}\left(\sigma_j^n,v_j^n\right)-\lambda\left[\mathcal{F}\left(\sigma_j^n\right)-\mathcal{F}\left(\sigma_{j-1}^n\right)\right]\\&\quad-\frac{\Delta t}{\delta}\mathcal{G}\left(\sigma_j^n,v_j^n,\sigma_j^{n+1},v_j^{n+1}\right)+M\Delta t\mathcal{H}\left(v_j^{n+1}\right).\end{aligned}\tag{4.13}$$

The properties of the entropy solutions of the system (2.1) will be derived from the corresponding properties of the finite difference solutions generated by the scheme (2.10). The convergence of the finite difference solutions is first established by a proper application of Helly's theorem, cf., e.g., [16].

LEMMA 4.3. *Suppose* (σ^0, v^0) *is the initial data which satisfies all the assumptions in* (2.12) *and let* $(\sigma^N, v^N)_\Delta$ *be the piecewise constant representation of the data generated by the scheme* (2.10). *Then, as the mesh parameters* Δx *and* Δt *tend to zero, there is a subsequence of* $(\sigma^N, v^N)_\Delta$, *which converges in* $(L^1_{loc}(\mathbb{R}\times\mathbb{R}))^2$ *to a pair of functions* (σ, v). *Furthermore,* $\sigma(\cdot,t), v(\cdot,t) \in BV$, *for all* $t \geq 0$, *and* $(\sigma(x,y), v(x,t)) \in \mathcal{S}$ *for* $(x,t) \in \mathbb{R}\times\mathbb{R}^+_0$, *and the following estimates hold:*

1. $(\sigma(x,t), v(x,t)) \in \mathcal{S}, \quad \forall (x,t) \in \mathbb{R}\times\mathbb{R}^+_0,$
2. $TV(\sigma(\cdot,t)) + TV(v(\cdot,t)) \leq TV(\sigma^0) + TV(v^0),$
3. $\|p(\cdot,t)\|_1 \leq M\delta,$
4. $\|\sigma(\cdot,t) - \sigma(\cdot,\tau)\|_1 + \|v(\cdot,t) - v(\cdot,\tau)\|_1 \leq M|t-\tau|,$
5. $Lip^+(\sigma(\cdot,t)) \leq MLip^+(\sigma^0), \quad Lip^+(v(\cdot,t)) \leq MLip^+(v^0), \quad \forall t \geq 0.$

Here, M *is a constant independent of* t *and* δ.

From the entropy inequality in (4.13), we derived that the limit solution is the entropy solution of (2.1).

LEMMA 4.4. *The limit solution* (σ, v) *constructed in Lemma* 4.3 *is the entropy solution of the system* (2.1), *which satisfies the following Kruzkov-type inequality:*

$$
\begin{aligned}
(4.14)\ \int_0^T \int_{\mathbb{R}} & [G(\sigma,v,k,q)(|\sigma-k| + |v-q|)\phi_t + |\sigma-k|\phi_x]\, dx\, dt \\
& + \int_{\mathbb{R}} G\left(\sigma^0, v^0, k, q\right)\left(\left|\sigma^0 - k\right| + \left|v^0 - q\right|\right)\phi(x,0)\, dx \\
& - \int_{\mathbb{R}} G(\sigma(x,T), v(x,T), k, q)(|\sigma(x,T)-k| + |v(x,T)-q|)\phi(x,T)\, dx \\
& + M\int_0^T \int_{\mathbb{R}} [|v-q| - (v-q)sgn\ (\sigma-k)]\,\phi\, dx\, dt \\
& \geq \frac{1}{\delta}\int_0^T \int_{\mathbb{R}} G(\sigma,v,k,q)R(\sigma,v)[sgn\ (\sigma-k) - sgn\ (v-q)]\phi\, dx\, dt.
\end{aligned}
$$

Here, $(k,q) \in \mathcal{S}$ *and* $\phi \in \mathcal{D}_+(T)$ *is any test function with compact support. We recall that the function* $G = G(\sigma,v,k,q)$ *is defined as*

$$
G(\sigma,v,k,q) = \frac{g(\sigma+v) - g(k+q)}{(\sigma+v) - (k+q)}.
$$

Proof. Let $\phi \in \mathcal{D}_+(T)$ be a test function with compact support. We multiply the inequality in (4.13) by $\phi(x_j, t_n)$, then sum over $0 \leq n \leq N-1$ and $j \in \mathcal{Z}$, and apply summation by parts with respect to n and j, and we obtain the following:

$$
\begin{aligned}
\Delta t \sum_{n=0}^{N-1} \Delta x \sum_{j\in\mathcal{Z}} & \left[\mathcal{E}\left(\sigma_j^{n+1}, v_j^{n+1}\right)\frac{\phi\left(x_j, t_{n+1}\right) - \phi\left(x_j,t_n\right)}{\Delta t}\right. \\
& \left. +\ \mathcal{F}\left(\sigma_j^n\right)\frac{\phi\left(x_{j+1},t_n\right) - \phi\left(x_j,t_n\right)}{\Delta x}\right] \\
& + \Delta x \sum_{j\in\mathcal{Z}} \mathcal{E}\left(\sigma_j^0, v_j^0\right)\phi\left(x_j, t^0\right) - \Delta x \sum_{j\in\mathcal{Z}} \mathcal{E}\left(\sigma_j^N, v_j^N\right)\phi\left(x_j, t^N\right)
\end{aligned}
$$

$$+\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathcal{Z}}M\mathcal{H}\left(v_j^{n+1}\right)\phi\left(x_j,t_n\right)$$
$$\geq\frac{1}{\delta}\Delta t\sum_{n=0}^{N-1}\Delta x\sum_{j\in\mathcal{Z}}\mathcal{G}\left(\sigma_j^n,v_j^n,\sigma_j^{n+1},v_j^{n+1}\right)\phi\left(x_j,t_n\right).$$

Now, by letting $\Delta x,\Delta t\to 0$ in the previous inequality, we get

$$\int_0^T\int_{\mathbb{R}}[\mathcal{E}(\sigma,v)\phi_t+\mathcal{F}(\sigma)\phi_x+M\mathcal{H}(v)\phi]\,dx\,dt$$
$$+\int_{\mathbb{R}}[\mathcal{E}(\sigma^0,v^0)\phi(x,0)-\mathcal{E}(\sigma(x,T),v(x,T))\phi(x,T)]\,dx$$
$$\geq\frac{1}{\delta}\int_0^T\int_{\mathbb{R}}\mathcal{G}(\sigma,v,\sigma,v)\phi\,dx\,dt.$$

Hence, by choosing a sequence of smooth function pairs $(\mathcal{E}_\theta,\mathcal{F}_\theta,\mathcal{G}_\theta,\mathcal{H}_\theta)$ such that, as $\theta\to 0$,

$$\mathcal{E}_\theta\to G(\sigma,v,k,q)(|\sigma-k|+|v-q|),$$
$$\mathcal{F}_\theta\to|\sigma-k|,$$
$$\mathcal{G}_\theta\to G(\sigma,v,k,q)R(\sigma,v)[\ \mathrm{sgn}\ (\sigma-k)-\ \mathrm{sgn}\ (v-q)],$$
$$\mathcal{H}_\theta\to|v-q|-(v-q)\ \mathrm{sgn}\ (\sigma-k),$$

uniformly, and we get the inequality (4.14) in Lemma 4.4 by the dominated convergence theorem. □

The uniqueness and continuous dependence with respect to the initial data in L^1 is then obtained by the Kruzkov-type argument.

LEMMA 4.5. *Let (σ,v) and $(\bar\sigma,\bar v)$ be two entropy solutions of the system* (2.1) *with initial data (σ^0,v^0) and $(\bar\sigma^0,\bar v^0)$, respectively. Then,*

$$\|\sigma(\cdot,t)-\bar\sigma(\cdot,t)\|_{L^1}+\|v(\cdot,t)-\bar v(\cdot,t)\|_{L^1}\leq\bar Me^{Mt}\left[\left\|\sigma^0-\bar\sigma^0\right\|_{L^1}+\left\|v^0-\bar v^0\right\|_{L^1}\right].$$

Proof. The uniqueness of the entropy solutions is proved by generalizing the arguments by Kruzkov [5] for scalar conservation laws. In this paper, only the sketch of the proof is given, and we refer to [14, 16] for the details in the proof.

For any $\theta\in(0,1]$, we introduce the mollifier function ω_θ on $\mathbb{R}$ as

$$\omega_\theta(x)=\frac{1}{\theta}\Omega\left(\frac{x}{\theta}\right),$$

where $\Omega:\mathbb{R}\to\mathbb{R}$ is a nonnegative, symmetric $\mathcal{C}^\infty$-function with support in $[-1,1]$ and satisfying

$$\int_{\mathbb{R}}\Omega(x)\,dx=1.$$

Let $T>0$. In (4.14), we choose $(k,q)=(\bar\sigma(y,\tau),\bar v(y,\tau))$ and $\phi(x,t)=\omega_\theta(x-y)\omega_\theta(t-\tau)$ for solution (σ,v), and integrate over $\mathbb{R}\times[0,T]$ with respect to y and τ, and we get an inequality. For the solution $(\bar\sigma,\bar v)$, we perform a similar operation, but

where we reverse the role of the variable (x,t) and (y,τ), we get another inequality. Now, adding these two inequalities, we get

$$L(\theta)+\frac{1}{\delta}l(\theta)\leq R(\theta)+2Mr(\theta),$$

where

$$\begin{aligned}L(\theta)=&\int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}}G(\sigma(x,T),v(x,T),\bar\sigma,\bar v)(|\sigma(x,T)-\bar\sigma|+|v(x,T)-\bar v|)\\&\qquad\omega_\theta(x-y)\omega_\theta(T-\tau)\,dx\,dy\,d\tau\\&+\int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}}G(\bar\sigma(y,T),\bar v(y,T),\sigma,v)(|\bar\sigma(y,T)-\sigma|+|\bar v(y,T)-v|)\\&\qquad\omega_\theta(x-y)\omega_\theta(T-\tau)\,dx\,dy\,d\tau\end{aligned}$$

and

$$\begin{aligned}R(\theta)=&\int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}}G(\sigma(x,0),v(x,0),\bar\sigma,\bar v)(|\sigma(x,0)-\bar\sigma|+|v(x,0)-\bar v|)\\&\qquad\omega_\theta(x-y)\omega_\theta(\tau)\,dx\,dy\,d\tau\\&+\int_0^T\int_{\mathbb{R}}\int_{\mathbb{R}}G(\bar\sigma(y,0),\bar v(y,0),\sigma,v)(|\bar\sigma(y,0)-\sigma|+|\bar v(y,0)-v|)\\&\qquad\omega_\theta(x-y)\omega_\theta(\tau)\,dx\,dy\,d\tau\end{aligned}$$

$$\begin{aligned}l(\theta)=&\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}}G(\sigma,v,\bar\sigma,\bar v)[\ \mathrm{sgn}\ (\sigma-\bar\sigma)-\ \mathrm{sgn}\ (v-\bar v)]\\&\qquad[R(\sigma,v)-R(\bar\sigma,\bar v)]\omega_\theta(x-y)\omega_\theta(t-\tau)\,dx\,dt\,dy\,d\tau\end{aligned}$$

and

$$r(\theta)=2\int_0^T\int_{\mathbb{R}}\int_0^T\int_{\mathbb{R}}|v-\bar v|\omega_\theta(x-y)\omega_\theta(t-\tau)\,dx\,dt\,dy\,d\tau.$$

First we note that $l(\theta)$ is non-negative. In order to estimate the turns $L(\theta)$ and $R(\theta)$, we introduce the function $\mathcal{N}(t)$ as

$$\mathcal{N}(t)=\int_{\mathbb{R}}G(\sigma(x,t),v(x,t),\bar\sigma(x,t),\bar v(x,t))(|\sigma(x,t)-\bar\sigma(x,t)|+|v(x,t)-\bar v(x,t)|)\,dx.$$

Note that the function $\mathcal{N}(t)$ is equivalent to

$$A(t):=\|\sigma(\cdot,t)-\bar\sigma(\cdot,t)\|_{L^1}+\|v(\cdot,t)-\bar v(\cdot,t)\|_{L^1}$$

in the sense that there exist two positive constants, M_1,M_2, such that

$$M_1A(t)\leq\mathcal{N}(t)\leq M_2A(t). \tag{4.15}$$

Then, as $\theta\to 0$, we get(cf., e.g., [16])

$$L(\theta)\to\mathcal{N}(T),\quad R(\theta)\to\mathcal{N}(0),$$

and

$$r(\theta) \to 2M \int_0^T \|v(\cdot,t) - \bar{v}(\cdot,t)\|_{L^1}\, dt.$$

Combining these estimates we conclude, in the limit case as $\theta \to 0$, that

$$\mathcal{N}(T) \le \mathcal{N}(0) + M \int_0^T \mathcal{N}(t)\, dt,$$

where M is a finite constant independent of δ. Thus, it follows that

$$\mathcal{N}(T) \le \mathcal{N}(0) e^{MT},$$

and again, using (4.15), we get

$$\|\sigma(\cdot,t) - \bar{\sigma}(\cdot,t)\|_{L^1} + \|v(\cdot,t) - \bar{v}(\cdot,t)\|_{L^1} \le \bar{M} e^{Mt} \left[\left\|\sigma^0 - \bar{\sigma}^0\right\|_{L^1} + \left\|v^0 - \bar{v}^0\right\|_{L^1} \right],$$

where $\bar{M}$ and M are finite constants independent of δ. This completes the proof of Theorem 2.2. ∎

5. Rate of convergence towards equilibrium: Proof of Theorem 2.3 and Corollary 2.4. We recall that Lemma 4.3 establishes bounds, uniformly with respect to δ, on the solutions $(\sigma_\delta, v_\delta)$ of the non-equilibrium model (1.1) or (2.1). By combining these estimates with standard compactness arguments we could have concluded, more or less directly, that these solutions converge to a solution of the equilibrium model (1.2) or (2.2) as the relaxation parameter δ tends to zero. However, we are not only interested in convergence, but also in a rate of convergence. Hence, in order to prove the error estimates in Theorem 2.3 and Corollary 2.4, we shall follow the work of Tadmor [15] and Kurganov and Tadmor [6]. First we observe that the entropy solutions of (1.1) are weak solutions of a scalar equation with an "error term."

LEMMA 5.1. *Let (u,σ) (resp., (σ,v)) be the entropy solutions of* (1.1) (*resp.,* (2.1)). *Then the solutions u are weak solutions of the following "error equation"*

$$u_t + \mu f(u)_x = -R(\sigma,v)_x$$

in the sense that the following integral equation holds for all test functions $\phi \in \mathcal{D}_+(T)$:

$$\int_0^T \int_{\mathbb{R}} (u\phi_t + \mu f(u)_x \phi_x)\, dx\, dt + \int_{\mathbb{R}} [u(x,0)\phi(x,0) - u(x,T)\phi(x,T)]\, dx = -\int_0^T \int_{\mathbb{R}} R(\sigma,v)\phi_x\, dx\, dt.$$

In addition, u satisfies the Lip$^+$ bound

$$\mathrm{Lip}^+(u(\cdot,t)) \le M, \quad \forall t \ge 0.$$

Proof. Let (σ,v) be the entropy solutions of (2.1). Then they satisfy the Kruzkov-type inequality given in (2.6). Choosing $(k = \sigma_m, q = v_m)$, where $\sigma_m = \min(\sigma)$ and $v_m = \min(v)$, (one can use, e.g., $k = q = 0$), the last terms on the left-hand side and the right-hand side are 0. Using the definition of G, the relation $u = g(\sigma + v)$, and the fact that (k,q) are constants, we get

$$\int_0^T \int_{\mathbb{R}} [u\phi_t + \sigma\phi_x]\, dx\, dt + \int_{\mathbb{R}} (u(x,0)\phi(x,0) - u(x,T)\phi(x,T)) \ge 0.$$

Similarly, by choosing $(k = \sigma_M, q = v_M)$, where $\sigma_M = \max(\sigma)$ and $v_M = \max(v)$ (e.g., $k = \mu, q = 1 - \mu$), we get

$$\int_0^T \int_{\mathbb{R}} [u\phi_t + \sigma\phi_x] \, dx \, dt + \int_{\mathbb{R}} (u(x,0)\phi(x,0) - u(x,T)\phi(x,T)) \leq 0.$$

These two inequalities lead to

$$\int_0^T \int_{\mathbb{R}} [u\phi_t + \sigma\phi_x] \, dx \, dt + \int_{\mathbb{R}} [u(x,0)\phi(x,0) - u(x,T)\phi(x,T)] = 0.$$

Furthermore, using the relation

$$\sigma - \mu f(u) = \sigma - \mu(\sigma + v) = (1 - \mu)\sigma - \mu v = R(\sigma, v),$$

we get the weak formulation in Lemma 5.1, and thus u is a weak solution of the error equation. Finally, the Lip^+ bound follows from the monotonicity of the function g. ∎

Let $T > 0$ be given and define $E = -R_x = -p_x$. Hence, $u = u_\delta$ is a weak solution of the inhomogeneous equation

$$u_t + \mu f(u)_x = E,$$

and $\bar{u}$ is a solution of the corresponding homogeneous equation (1.2). Furthermore, these solutions satisfy an Oleinik condition of the form

$$\mathrm{Lip}^+(u(\cdot,t)),\ \mathrm{Lip}^+(\bar{u}(\cdot,t) \leq M, \quad \forall t \geq 0.$$

Since the flux function f is convex, we can therefore conclude from the arguments in Kurganov and Tadmor [6] that the following stability estimate holds:

$$\|u(\cdot,t) - \bar{u}(\cdot,t)\|_{\mathrm{Lip}'} \leq M \sup_{0 \leq \tau \leq t} \|E(\cdot,\tau)\|_{\mathrm{Lip}'}, \quad 0 \leq t \leq T.$$

From Lemma 4.3 we obtain that

$$\|E(\cdot,t)\|_{\mathrm{Lip}'} \leq \|p(\cdot,t)\|_{L^1} \leq M\delta.$$

This completes the proof of Theorem 2.3.

The L^p estimate in Corollary 2.4 can be proved by interpolation between the Lip'-error estimate in Theorem 2.3 and the BV-boundness of the error, exactly in the same way as is done in Nessyahu and Tadmor [11]. We therefore omit the details.

The L^1 estimate for $\sigma - \bar{\sigma}$ follows from the L^1 estimate for $u - \bar{u}$. To be precise, since $\bar{\sigma} = \mu f(\bar{u})$, we have

$$\begin{aligned}
\|\sigma - \bar{\sigma}\|_{L^1} &= \|\sigma - \mu f(u) + \mu f(u) - \bar{\sigma}\|_{L^1} \leq \|\sigma - \mu f(u)\|_{L^1} + \|\mu(f(u) - f(\bar{u}))\|_{L^1} \\
&\leq \|p\|_{L^1} + M\|u(\cdot,t) - \bar{u}(\cdot,t)\|_{L^1} \\
&\leq M\sqrt{\delta},
\end{aligned}$$

which gives the second estimate in Corollary 2.4.

Acknowledgments. The authors are grateful to one of the referees for pointing out possible ways of improving the results in an earlier version of this paper. We also thank Professor Natalini for informing us about the works of Yong [18] and Luo and Natalini [9].

REFERENCES

[1] G. Q. Chen and T. P. Liu, *Zero relaxation and dissipation limits for hyperbolic conservation laws*, Comm. Pure Appl. Math., XLVI (1993), pp. 755–781.

[2] G. Q. Chen, C. D. Levermore, and T. P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[3] K. N. Chueh, C. C. Conley, and J. A. Smoller, *Positively invariant regions for systems of nonlinear diffusion equations*, Indiana Univ. Math. J., 26 (1977), pp. 373–392.

[4] J. M. Greenberg and L. Hsiao, *The Riemann problem for the system* $u_t + \sigma_x = 0$ *and* $(\sigma - f(u))_t + (\sigma - \mu f(u)) = 0$, Arch. Rational Mech. Anal., 82 (1983), pp. 87–108.

[5] S. N. Kruzkov, *First order quasi linear equations with several space variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[6] A. Kurganov and E. Tadmor, *Stiff Systems of Hyperbolic Conservation Laws. Convergence and Error Estimates*, preprint, UCLA, Los Angeles, April 1996.

[7] N. N. Kuznetsov, *The accuracy of certain approximate methods for the computation of weak solutions of a first order quasilinear equation*, Comput. Math. Math. Phys., 16 (1976), pp. 105–119 (translation).

[8] T. P. Liu, *Hyperbolic conservation laws with relaxation*, Comm. Math. Phys., 108 (1987), pp. 153–175.

[9] T. Luo and R. Natalini, *BV solutions and relaxation limit for a model in viscoelasticity*, Proc. Roy. Soc. Edinburgh, 128A (1998), pp. 775–795.

[10] R. Natalini, *Convergence to equilibrium for the relaxation approximations of conservation laws*, Comm. Pure Appl. Math., 49 (1996), pp. 1279–1292.

[11] H. Nessyahu and E. Tadmor, *The convergence rate of approximate solutions for nonlinear scalar conservation laws*, SIAM J. Numer. Anal., 29 (1992), pp. 1505–1519.

[12] O. Oleinik, *Discontinuous solutions of nonlinear differential equations*, Amer. Math. Soc. Transl. Ser. 2, 26 (1963), pp. 95–172.

[13] H. J. Schroll, A. Tveito, and R. Winther, *An* L^1*-error bound for a semi-implicit difference scheme applied to a stiff system of conservation laws*, SIAM J. Numer. Anal., 34 (1997), pp. 1152–1166.

[14] W. Shen, A. Tveito, and R. Winther, *A system of conservation laws including a stiff relaxation term: the* $2D$ *case*, BIT, 36 (1996), pp. 786–813.

[15] E. Tadmor, *Local error estimates for discontinuous solutions of nonlinear hyperbolic equations*, SIAM J. Numer. Anal., 28 (1991), pp. 891–906.

[16] A. Tveito and R. Winther, *On the rate of convergence to equilibrium for a system of conservation laws including a relaxation term*, SIAM J. Math. Anal, 28 (1997), pp. 136–161.

[17] G. B. Whitham, *Linear and Non-Linear Waves*, Wiley, New York, 1974.

[18] W.-A. Yong, *A Difference Scheme for a Stiff System of Conservation Laws*, Preprint 95-25 (SFB 359), Interdisziplinares Zentrum fur wissenschaftliches Rechnen, University of Heidelberg, Germany, 1995.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1136–1150

ALMOST EVERYWHERE SOLUTIONS OF PARTIAL DIFFERENTIAL EQUATIONS AND SYSTEMS OF ANY ORDER*

LAURA POGGIOLINI†

Abstract. We prove existence of $W^{k,\infty}(\Omega)$ solutions to some differential problems related to nonlinear partial differential equations and systems of order $k \geq 1$. The results will be proved by means of Baire's lemma.

Key words. quasi-convexity, almost everywhere solutions, systems of PDEs

AMS subject classification. 35G30

PII. S0036141098336121

1. Introduction.

In this paper we deal with the existence of almost everywhere (a.e.) solutions to differential problems of the following kind:

$$\begin{cases} F(x, D^{[k-1]}u(x), D^k u(x)) = 0, & \text{a.e. } x \in \Omega, \\ u \in \phi + W_0^{k,\infty}(\Omega), \end{cases} \tag{1.1}$$

where k is an integer greater than or equal to 1 (for notations and definitions see the next section) and F is a given function which must satisfy certain coercivity conditions; in particular our hypotheses will rule out functions F which are linear with respect to the higher-order derivatives $D^k u$. We shall prove the following theorem.

THEOREM 1.1. *Let $\Omega \subset \mathbb{R}^n$ be an open Lipschitz set. Let $F\colon \Omega \times \cdots \times ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$ be a continuous function, quasi-convex and coercive in a rank-1 direction (see Definition 2.3) with respect to the last variable. Let $\phi\colon \overline{\Omega} \to \mathbb{R}$ be a $C^k(\overline{\Omega})$ function such that*

$$F(x, D^{[k-1]}\phi(x), D^k\phi(x)) \leq 0 \qquad \forall x \in \overline{\Omega}.$$

Then there exists a function $u \in \phi + W_0^{k,\infty}(\Omega)$ such that

$$F(x, D^{[k-1]}u(x), D^k u(x)) = 0, \qquad \text{a.e. } x \in \Omega. \tag{1.2}$$

As we already said, the definition of quasi-convexity is discussed in section 2. Note however that quasi-convexity in the context of Theorem 1.1 is equivalent to the usual convexity in the first-order case $k = 1$.

Theorem 1.1 extends to the higher-order case $k \in \mathbb{N}$ an analogous result recently proved by Dacorogna and Marcellini in the first-order case $k = 1$ (see [7], [6], [8] and [9]) and in the second-order case $k = 2$ (see [10]). We will use the same techniques based on Baire category method, exploited by Dacorogna and Marcellini to study vector valued problems and originally introduced by Cellina in [3] in the context of ordinary differential inclusions (see also De Blasi and Pianigiani [11], [12] and Bressan and Flores [2]).

We should refer also to some previous research for the first- and second-order cases in the context of viscosity solutions (see, for example, Crandall and Lions [4];

*Received by the editors March 20, 1998; accepted for publication (in revised form) November 2, 1998; published electronically August 26, 1999.
http://www.siam.org/journals/sima/30-5/33612.html
†Dipartimento di Matematica "U. Dini," Viale Morgagni 67/a, 50134 Firenze, Italy (laura.poggiolini@math.unifi.it).

see also [10] for some other references). We should also mention the recent works by Müller and Svërak, [21] and [22] using the method of convex integration introduced by Gromov [14].

A main difficulty in considering $k \geq 2$ instead of $k = 1$ is the loss of approximating a given $W^{k,\infty}(\Omega)$ function u by a sequence of piecewise polynomials of degree k. In our context of higher-order $k \in \mathbb{N}$, the main tool to apply Baire category method is the density argument obtained in Lemma 3.1. Other parts are treated in similar ways to [10].

If we consider for a moment the case $k = 4$ and $F = F(D^4u(x))$, it is evident that we can't deal with a linear function F; in fact the problem

$$\begin{cases} \Delta^2 u(x) - 1 = 0, & \text{a.e. } x \in \Omega, \\ u \in \phi + W_0^{4,\infty}(\Omega) \end{cases} \tag{1.3}$$

(here Δ^2 denotes the bilaplacian operator) is over determined, while we shall provide a solution, for example, to the problem

$$\begin{cases} |\Delta^2 u(x)| - 1 = 0, & \text{a.e. } x \in \Omega, \\ u \in \phi + W_0^{4,\infty}(\Omega). \end{cases} \tag{1.4}$$

Actually we shall deal only with coercive functions F (see Definition 2.3), such as $F(D^4u(x)) = |\Delta^2 u(x)|$ in the previous example (1.4). In section 6 we shall deal more generally with k-order systems of the type

$$\begin{cases} F_i(x, D^{[k-1]}u(x), D^k u(x)) = 0, & \text{a.e. } x \in \Omega, \quad i = 1, \dots, N, \\ u \in \phi + W_0^{k,\infty}(\Omega) \end{cases}$$

and we will prove an existence theorem for such systems.

2. Notation and definitions. If $v \in \mathbb{R}^n$ and $k \in \mathbb{N}$ we denote by $v^{\otimes_k}$ the tensor product $\overbrace{v \otimes \cdots \otimes v}^{k\text{-times}}$. The symbol $((\mathbb{R}^n)^{\otimes_k})_s$ will denote the subset of symmetric tensors of the space $(\mathbb{R}^n)^{\otimes_k} \equiv \underbrace{\mathbb{R}^n \otimes \cdots \otimes \mathbb{R}^n}_{k\text{-times}}$.

If $\Omega \subset \mathbb{R}^n$ is a Lebesgue measurable set, the symbol $|\Omega|$ will denote its n-dimensional Lebesgue measure. If $u\colon \Omega \subset \mathbb{R}^n \to \mathbb{R}$ is a weakly k-differentiable function we denote $D^k u = (D^\alpha u)_{|\alpha|=k}$ and $D^{[k]}u = (u, Du, \dots, D^k u)$.

If ψ is a $C^0(\overline{\Omega})$ function, let $\|\psi\|_{\infty,\Omega} \equiv \sup_{x\in\Omega} |\psi(x)|$, while if $\phi \in C^l(\overline{\Omega})$, let $\|\phi\|_{l,\infty,\Omega} \equiv \sum_{i=0}^{l} \sup_{x\in\Omega} |D^i\psi(x)|$.

The following definitions of convexity are known extensions for the higher-order case $k > 1$ of analogous convexity conditions for the first-order vector valued case (see Morrey [19], [20], Meyers [17], Dacorogna [5], and Ball, Currie, and Olver [1]).

DEFINITION 2.1. *We say that $\Lambda \in ((\mathbb{R}^n)^{\otimes_k})_s$ is a rank-1 tensor if there exist $\mu \in \mathbb{R}$, $v \in \mathbb{R}^n$ such that*

$$\Lambda = \mu v^{\otimes_k}. \tag{2.1}$$

We remark that in Definition 2.1 we can always assume that the Euclidean norm of v, $|v|$ is 1.

DEFINITION 2.2. *We say that a function $F\colon ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$ is quasi-convex if*

$$\int_\Omega F(A + D^k\phi(x))\,dx \geq |\Omega| F(A) \quad \forall A \in ((\mathbb{R}^n)^{\otimes_k})_s \tag{2.2}$$

$$\forall \Omega \subset \mathbb{R}^n \text{ open bounded set and } \forall \phi \in C_0^\infty(\Omega).$$

DEFINITION 2.3. *Let Y be a metric space. We say that a function $F: Y \times ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$ is coercive in a rank-1 direction Λ if for any bounded subset B of $Y \times ((\mathbb{R}^n)^{\otimes_k})_s$ there exist m, $q \in \mathbb{R}$, with $m > 0$ such that*

$$F(y, \xi + t\Lambda) \geq m|t| - q \qquad \forall t \in \mathbb{R},\ \forall (y, \xi) \in B. \tag{2.3}$$

DEFINITION 2.4. *We denote $P_\varepsilon(\Omega)$ the class of the functions $u \in W^{k,\infty}(\Omega)$ such that there exist Ω_0, Ω_j, $j \in \mathbb{N}$ open Lipschitz pairwise disjoint subsets of Ω such that*

$$\begin{aligned} &|\Omega_0| < \varepsilon, \\ &\bigcup_{j \geq 0} \overline{\Omega_j} = \overline{\Omega}, \\ &D^k u = \xi_j, \qquad \text{a.e. } x \in \Omega_j \quad \forall j \in \mathbb{N}. \end{aligned}$$

Most of the proofs will be carried out by means of Baire's lemma (see, e.g., Kolmogorov and Fomin [16]).

LEMMA 2.5 (Baire's lemma). *If V is a complete metric space and V^m, $m \in \mathbb{N}$ are open dense subsets of V, then also $\bigcap_{m \in \mathbb{N}} V^m$ is dense in V.*

3. The main approximation lemma. We begin with a fundamental technical lemma.

LEMMA 3.1. *Let $\Omega \subset \mathbb{R}^n$ be an open Lipschitz set. Let $t \in [0,1]$ and let A, $B \in ((\mathbb{R}^n)^{\otimes_k})_s$ be such that $\operatorname{rank}\{A - B\} = 1$. Let $\phi: \Omega \to \mathbb{R}$ be such that $D^k \phi(x) = tA + (1-t)B = B + t(A - B)$ for every $x \in \Omega$. Then for any $\varepsilon > 0$ there exist a function $u \in \phi + W_0^{k,\infty}(\Omega)$ and two open Lipschitz disjoint subsets of Ω, Ω_A and Ω_B, such that*

$$\|\Omega_A| - t|\Omega\| \leq \varepsilon, \quad ||\Omega_A| - (1-t)|\Omega|| \leq \varepsilon, \tag{3.1}$$

$$\|u - \phi\|_{k-1,\infty} < \varepsilon, \tag{3.2}$$

$$D^k u(x) = \begin{cases} A, & x \in \Omega_A, \\ B, & x \in \Omega_B, \end{cases} \tag{3.3}$$

$$\operatorname{dist}(D^k u(x),\ co(A, B)) \leq \varepsilon, \qquad \text{a.e. } x \in \Omega. \tag{3.4}$$

Proof. We divide the proof into several steps.

1. Let us assume that $A - B = \mu e_1^{\otimes_k}$, where $e_1 = (1, 0, \ldots, 0)$ is the first vector of the canonical basis of $\mathbb{R}^n$.

2. We can write Ω as the disjoint union of cubes whose faces are parallel to the coordinates axes and of a set of small measure. If we define $u = \phi$ in this set, then up to homotheties and translations we can always assume that $\Omega = (0,1)^n$.

3. Let $\Omega_\varepsilon \subset\subset \Omega$ be such that $|\Omega \setminus \Omega_\varepsilon| < \varepsilon$ and let $\eta \in C_0^k(\Omega)$ be such that

$$\begin{aligned} &0 \leq \eta(x) \leq 1 \qquad \forall x \in \Omega, \\ &\eta(x) = 1 \qquad \forall x \in \Omega_\varepsilon, \\ &|D^l \eta(x)| \leq \frac{L}{\varepsilon^l} \qquad \forall l = 1, \ldots, k \ \text{ and } \ \forall x \in \Omega \setminus \Omega_\varepsilon. \end{aligned}$$

We now define a new function $v\colon (0,1) \to \mathbb{R} \quad v\colon x_1 \to v(x_1)$. Let $\delta > 0$, $x_1 \in (0,1)$ and let I, $J \subset (0,1)$ be such that I and J are the disjoint union of open disjoint intervals and

$$\begin{aligned}
&I \cap J = \emptyset, \quad \overline{I} \cup \overline{J} = [0,1],\\
&|I| = t, \quad |J| = 1 - t,\\
&v^{(k)}(x_1) = \begin{cases} (1-t)\mu x_1 \in I,\\ -t\mu x_1 \in J,\end{cases}\\
&|v^l(x_1)| < \delta \qquad \forall x_1 \in (0,1) \text{ and } \forall l = 1, \dots, k.
\end{aligned}$$

Now let us define $u(x) = u(x_1, x') = \phi(x) + \eta(x)v(x_1) = \eta(x)(v(x_1) + \phi(x)) + (1 - \eta(x))\phi(x)$ and set

$$\begin{aligned}
\Omega_A &\equiv \{x \in \Omega_\varepsilon\colon x_1 \in I\},\\
\Omega_B &\equiv \{x \in \Omega_\varepsilon\colon x_1 \in J\}.
\end{aligned}$$

Then the function u satisfies all the requests of our thesis:

$$D_{i_1,\dots,i_n} u = D_{i_1,\dots,i_n}\phi + \sum_{\substack{l_s + m_s = i_s\\ s=1,\dots,n}} D_{l_1,\dots,l_n}\eta D_{m_1,\dots,m_n} v.$$

If $x \in \Omega_A$ we have

$$\begin{aligned}
&D_{l_1,\dots,l_n}\eta = 0 \text{ whenever } l_1 + \cdots + l_n > 0,\\
&D_{m_1,\dots,m_n} v = (1-t)\mu\delta_{1m_1}, \dots, \delta_{1m_n} \text{ whenever } m_1 + \cdots + m_n = k,
\end{aligned}$$

hence,

$$D^k u = D^k\phi + (1-t)\mu e_1^{\otimes k} = B + t(A-B) + (1-t)(A-B) = A,$$

while, if $x \in \Omega_B$ we have

$$\begin{aligned}
&D_{l_1,\dots,l_n}\eta = 0 \text{ whenever } l_1 + \cdots + l_n > 0,\\
&D_{m_1,\dots,m_n} v = -t\mu\delta_{1m_1}, \dots, \delta_{1m_n}, \text{ whenever } m_1 + \cdots + m_n = k,
\end{aligned}$$

hence,

$$D^k u = D^k\phi - t\mu e_1^{\otimes k} = B + t(A-B) - t(A-B) = B.$$

To obtain (3.3) we only need to choose $\delta < C(n,k)\min\{1, \varepsilon, \dots, \varepsilon^k\}$.

Now let us prove inequality (3.4):

$$D^k\phi(x) = tA + (1-t)B \in co(A,B),$$

$$D^k\phi + D^k v = \begin{cases} tA + (1-t)B + (1-t)(A-B) = A, & x_1 \in I,\\ tA + (1-t)B - t(A-B) = B, & x_1 \in J,\end{cases}$$

which belong to $co(A,B)$. But

$$\begin{aligned}
D_{i_1,\dots,i_n} u &= D_{i_1,\dots,i_n}\phi + \sum_{\substack{l_s + m_s = i_s\\ s=1,\dots,n}} D_{l_1,\dots,l_n}\eta D_{m_1,\dots,m_n} v\\
&= \eta D_{i_1,\dots,i_n}\phi + (1-\eta) D_{i_1,\dots,i_n}\phi + \eta D_{i_1,\dots,i_n} v\\
&\quad + \sum_{\substack{l_1 + \cdots + l_n > 0\\ l_s + m_s = i_s}} D_{l_1,\dots,l_n}\eta D_{m_1,\dots,m_n} v,
\end{aligned}$$

which means

$$D^k u = \eta(D^k\phi + D^k v) + (1-\eta)D^k\phi + P_k$$

where

$$(P_k)_{i_1,\dots,i_n} = \sum_{\substack{l_s+m_s=i_s \\ l_1+\cdots+l_s>0}} D_{l_1,\dots,l_s}\eta D_{m_1,\dots,m_n}v.$$

Therefore

$$\operatorname{dist}(D^k u, co(A,B)) \le \|P_k\|_\infty \le C(n,k)L\max\{\varepsilon^{-1},\dots,\varepsilon^{-k+1}\}\delta.$$

We choose

$$\delta \le (C(n,k))^{-1}\min\{\varepsilon^2,\dots,\varepsilon^k\}$$

and obtain inequality (3.4).

4. Let us see why Lemma 3.1 holds also when $v \neq e_1$. Let us define a matrix R in the following way. We define $r_{i1} = v^i\ \forall i = 1,\dots,n$. We can choose the other elements of R in such a way that R is an orthogonal matrix and $v = Re_1$, hence $A - B = \mu(Re_1)^{\otimes_k}$. If $\phi\colon \overline{\Omega} \subset \mathbb{R}^n \to \mathbb{R}$ is a C^k function, we define $\widetilde{\phi}\colon R^t\Omega \to \mathbb{R}$ $\widetilde{\phi}\colon y \to \phi(Ry)$. If we compute the partial derivatives of order k we have

$$\frac{\partial^k\widetilde{\phi}}{\partial y_{j_1},\dots,\partial y_{j_k}}(y) = \sum_{i_1,\dots,i_k=1}^{n} \frac{\partial^k\phi}{\partial x_{i_1},\dots,\partial x_{i_k}}(Ry)r_{i_1j_1},\dots,r_{i_kj_k}.$$

But, since $D^k\phi(x) = \mu v^{\otimes_k} = \mu(Re_1)^{\otimes_k}$, we get

$$\left(D^k\phi\right)_{i_1,\dots,i_k}(x) = \mu r_{i_1 1},\dots,r_{i_k 1}.$$

Therefore

$$\begin{aligned}\frac{\partial^k\widetilde{\phi}}{\partial y_{j_1},\dots,\partial y_{j_k}}(y) &= \sum_{i_1,\dots,i_k=1}^{n} \mu r_{i_1 1},\dots,r_{i_k 1}r_{i_1j_1},\dots,r_{i_kj_k} \\ &= \sum_{i_1,\dots,i_k=1}^{n} \mu(R^t)_{1i_1},\dots,(R^t)_{1i_k}(R)_{i_1j_1},\dots,(R)_{i_kj_k} \\ &= \mu\delta_{1j_1},\dots,\delta_{1j_k} = \mu(e_1^{\otimes_k})_{j_1,\dots,j_k},\end{aligned}$$

which means

$$D^k\widetilde{\phi}(y) = \mu(e_1)^{\otimes_k}.$$

So, applying the previous steps, we can find a function $\widetilde{u}$ and two open disjoint Lipschitz subsets $\widetilde{\Omega}_{\widetilde{A}}$ and $\widetilde{\Omega}_{\widetilde{B}}$ that solve the problem in $\widetilde{\Omega}$ for the datum $\widetilde{\phi}$ and the tensors $\widetilde{A}$ and $\widetilde{B}$ defined as follows:

$$\begin{aligned}(\widetilde{A})_{j_1,\dots,j_k} &= \sum_{i_1,\dots,i_k=1}^{n} (A)_{i_1,\dots,i_k}r_{i_1j_1},\dots,r_{i_kj_k}, \\ (\widetilde{B})_{j_1,\dots,j_k} &= \sum_{i_1,\dots,i_k=1}^{n} (B)_{i_1,\dots,i_k}r_{i_1j_1},\dots,r_{i_kj_k}.\end{aligned}$$

The function $u\colon \Omega \to \mathbb{R}$ $u\colon x \to \widetilde{u}(R^t x)$ and the sets $\Omega_A \equiv R\widetilde{\Omega}_{\widetilde{A}}$ and $\Omega_B \equiv R\widetilde{\Omega}_{\widetilde{B}}$ will satisfy the thesis of the lemma. □

4. A model case. Now we give a first existence result which holds when F depends only on the highest-order derivatives and when the boundary datum ϕ is a polynomial.

THEOREM 4.1. *Let $\Omega \subset \mathbb{R}^n$ be an open Lipschitz set. Let $F\colon ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$ be a quasi-convex function, coercive in a rank-1 direction Λ. Let $\phi\colon \overline{\Omega} \to \mathbb{R}$ be a polynomial of degree less than or equal to k such that $F(D^k\phi(x)) = F(\xi_0) \le 0$. Then there exists a function $u \in \phi + W_0^{k,\infty}(\Omega)$ such that $F(D^k u(x)) = 0$, a.e. $x \in \Omega$.*

Proof. We can assume that Ω is bounded and that $F(\xi_0) < 0$, otherwise $u = \phi$ is a solution to our problem.

For $r > 0$ we define

$$K = \left\{\eta \in ((\mathbb{R}^n)^{\otimes_k})_s\colon \eta = \xi + t\Lambda, \quad \xi \in \overline{B(\xi_0, r)}, \quad t \in \mathbb{R}\colon m|t| - q \le 0\right\}.$$

K is a convex compact set and

$$\left\{\eta \in ((\mathbb{R}^n)^{\otimes_k})_s\colon \eta = \xi + t\Lambda, \quad \xi \in \overline{B(\xi_0, r)}, \quad F(\eta) \le 0\right\} \subset K.$$

Let

$$\begin{aligned} V = \{u \in \phi + W_0^{k,\infty}(\Omega)\colon \quad & \exists \varepsilon_l \downarrow 0 \exists u_l \in P_{\varepsilon_l} \text{ such that} \\ & \|u - u_l\|_{k-1,\infty,\Omega} \le \varepsilon, \\ & D^k u_l(x) \in \operatorname{int} K, \text{ a.e. } x \in \Omega, \\ & F(D^k u_l(x)) < 0, \text{ a.e. } x \in \Omega, \\ & u_l \in u + W_0^{k,\infty}(\Omega)\}. \end{aligned}$$

Since ϕ is in V, V is nonempty and $(V, C^{k-1}(\overline{\Omega}))$ is a complete metric space. But K is bounded, hence V is bounded in $W_0^{k,\infty}(\Omega)$: Let $u \in V$ and let u_l be a sequence approximating u; since $\|D^k u_l\|_\infty \le C$ there exists a function $g \in L^\infty(\Omega, ((\mathbb{R}^n)^{\otimes_k})_s)$ such that $D^k u_l$ converges (up to a subsequence) to g in the weak*-topology of L^∞. We want to prove that $g = D^k u_l$ a.e. Let g_s be one of the components of g (s is a multi-index). Let $s = (s_1, s')$, $s_1 \in \mathbb{N}$ and let $\psi \in C_0^1(\Omega)$. Then

$$\begin{aligned} \int_\Omega \psi g_s \, dx = \lim_{l\to\infty} \int_\Omega \psi D_s u_l \, dx = \lim_{l\to\infty} - \int_\Omega \left\langle \frac{\partial \psi}{\partial x_{s_1}}, D_{s'} u_l \right\rangle dx \\ - \int_\Omega \left\langle \frac{\partial \psi}{\partial x_{s_1}}, D_{s'} u \right\rangle dx = \int_\Omega \psi D_s u \, dx, \end{aligned}$$

which implies $g(x) = D^k u(x)$, a.e. $x \in \Omega$.

Therefore u_l converges to u in the weak*topology of $W_0^{k,\infty}(\Omega)$. Let $\eta \in C_0^\infty(\Omega)$ be a nonnegative function; by the quasi-convexity of F (see Meyers [18], Fusco [13], and Guidorzi and Poggiolini [15]) we have

$$0 \ge \liminf_{l\to\infty} \int_\Omega F(D^k u_l(x))\eta(x) \, dx \ge \int_\Omega F(D^k u(x))\eta(x) \, dx,$$

which implies

$$F(D^k u(x)) \le 0, \qquad \text{a.e. } x \in \Omega$$

and therefore

$$V \subset \{u \in \phi + W_0^{k,\infty}(\Omega)\colon F(D^k u(x)) \le 0, \text{ a.e. } x \in \Omega\}.$$

For each $m \in \mathbb{N}$ let us consider the set

$$V^m = \left\{ u \in V \colon \int_\Omega F(D^k u(x)) \geq \frac{-1}{m} \right\}.$$

V^m is an open subset of V, indeed by quasi-convexity of F, $V \setminus V^m$ is closed (see again [18], [13], [15]).

We want to prove that V^m is also a dense subset of $(V, C^{k-1}(\Omega))$. Let $v \in V$; by definition there exist $\varepsilon > 0$ and v_ε such that

$$\begin{aligned}
&v_\varepsilon \in P_\varepsilon(\Omega),\\
&D^k v_\varepsilon(x) \in \operatorname{int} K, \ \text{ a.e. } x \in \Omega,\\
&F(D^k v_\varepsilon(x)) < 0, \ \text{ a.e. } x \in \Omega,\\
&v_\varepsilon \in \phi + W_0^{k,\infty}(\Omega),\\
&\|v_\varepsilon - v\|_{k-1,\infty,\Omega} < \frac{\varepsilon}{2}.
\end{aligned}$$

These mean that there exist Ω_0, Ω_j, $j \in \mathbb{N}$ open Lipschitz pairwise disjoint subsets of Ω such that

$$\begin{aligned}
&\bigcup_{j \geq 0} \overline{\Omega_j} = \overline{\Omega},\\
&|\Omega_0| \leq \varepsilon,\\
&D^k v_\varepsilon(x) = \xi_j \in \operatorname{int} K, \ \text{ a.e. } x \in \Omega,\\
&F(\xi_j) < 0 \quad \forall j \in N.
\end{aligned}$$

For each $j \in \mathbb{N}$ let us consider the application $\tau_j \colon t \in \mathbb{R} \to \xi_j + t\Lambda \in ((\mathbb{R}^n)^{\otimes_k})_s$. Since $F(\xi_j + t\Lambda) \geq m|t| - q \quad \forall t \in \mathbb{R}$ and $F(\xi_j) < 0$, there exist t_1, t_2, $t_1 < 0 < t_2$, such that

$$F(\tau_j(t_1)) = F(\tau_j(t_2)) = 0$$

and therefore, there exist $\widetilde{t_1}$, $\widetilde{t_2}$, $t_1 < \widetilde{t_1} < 0 < \widetilde{t_2} < t_2$, such that

$$F(\tau_j(\widetilde{t_1})) > -\varepsilon, \qquad F(\tau_j(\widetilde{t_2})) > -\varepsilon, \tag{4.1}$$

$$\tau_j(\widetilde{t_1}) \in \operatorname{int} K, \qquad \tau_j(\widetilde{t_2}) \in \operatorname{int} K. \tag{4.2}$$

Moreover, since $\tau_j(\widetilde{t_2}) - \tau_j(\widetilde{t_1}) = (\widetilde{t_2} - \widetilde{t_1})\Lambda$ is a rank-1 tensor, we can apply Lemma 3.1 in the open Lipschitz set Ω_j to the tensors $A_j = \tau_j(\widetilde{t_1})$ and $B_j = \tau_j(\widetilde{t_2})$ with the boundary value v_ε.

There exist $\Omega_j^1, \Omega_j^2 \subset \Omega_j$, and a function v_e^j such that

$$\begin{aligned}
&|\Omega_j \setminus (\Omega_j^1 \cup \Omega_j^2)| < \varepsilon 2^{-j},\\
&v_\varepsilon^j \in v_\varepsilon + W_0^{k,\infty}(\Omega_j),\\
&D^k v_\varepsilon^j(x) = \xi_j + \widetilde{t_1}\Lambda, \ \text{ a.e. } x \in \Omega_j^1,\\
&D^k v_\varepsilon^j(x) = \xi_j + \widetilde{t_2}\Lambda, \ \text{ a.e. } x \in \Omega_j^2,\\
&\|v_\varepsilon^j - v_\varepsilon\|_{k-1,\infty,\Omega_j} < \varepsilon 2^{-j},\\
&D^k(v_\varepsilon^j(x)) \in \operatorname{int} K, \ \text{ a.e. } x \in \Omega_j,\\
&F(D^k(v_\varepsilon^j(x))) < 0, \ \text{ a.e. } x \in \Omega_j.
\end{aligned}$$

If we define

$$u_\varepsilon(x) = \begin{cases} v_\varepsilon(x), & x \in \Omega_0, \\ v_\varepsilon^j(x), & x \in \Omega_j, \quad j \in \mathbb{N}, \end{cases}$$

we have $u_\varepsilon \in P_{2\varepsilon}(\Omega)$, $\|u_\varepsilon - v_\varepsilon\|_{k-1,\infty,\Omega} < \varepsilon$; we need to show that $u_\varepsilon \in V^m$:

$$\begin{aligned} \int_\Omega F(D^k u_\varepsilon(x))\,dx &= \int_{\Omega_0} F(D^k v_\varepsilon(x))dx \\ &\quad + \sum_{j\in\mathbb{N}} \left[\int_{\Omega_j^1 \cup \Omega_j^2} F(D^k v_\varepsilon^j(x))\,dx + \int_{\Omega_j \setminus (\Omega_j^1 \cup \Omega_j^2)} F(D^k v_\varepsilon^j(x))\,dx \right] \\ &\geq -|\Omega_0|C - \sum_{j\in\mathbb{N}} \left[\varepsilon |\Omega_j \setminus (\Omega_j^1 \cup \Omega_j^2)| - C\varepsilon 2^{-j} \right] \\ &\geq -\varepsilon\,[C + |\Omega| + C] \geq \frac{-1}{m} \end{aligned}$$

for $\varepsilon > 0$ sufficiently small. To conclude our proof we only must apply Baire's lemma: the subsets V^m are open and dense in the complete metric space V, therefore $\bigcap_{m\in\mathbb{N}} V^m$ is also dense in V. Let $u \in \bigcap_{m\in\mathbb{N}} V^m$:

$$\begin{aligned} &F(D^k u(x)) \leq 0, \ \text{ a.e. } x \in \Omega \ \text{ because } \ u \in V; \\ &\int_\Omega F(D^k u(x))\,dx \geq 0 \ \text{ because } \ u \in \bigcap_{m\in\mathbb{N}} V^m; \end{aligned} \tag{4.3}$$

so it must be

$$F(D^k u(x)) = 0, \text{ a.e. } x \in \Omega. \qquad \square$$

5. Equation with lower-order terms. In this section we prove our main theorem concerning the case of only one equation, i.e., Theorem 1.1.

LEMMA 5.1. *Let $\Omega \subset \mathbb{R}^n$ be an open Lipschitz set. Let $F\colon \Omega \times \cdots \times ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$ be a continuous function, quasi-convex and coercive in a rank-1 direction Λ with respect to the last variable. Let $\phi\colon \widetilde{\Omega} \to \mathbb{R}$ be a $C^k(\overline{\Omega})$ function such that $F(x, D^{[k-1]}\phi(x), D^k\phi(x)) < 0$ $\forall x \in \overline{\Omega}$.*

Then there exists a function $u \in \phi + W_0^{k,\infty}(\Omega)$ such that

$$F(x, D^{[k-1]}u(x), D^k u(x)) = 0, \ \textit{a.e. } x \in \Omega. \tag{5.1}$$

Proof. As in the previous lemmas we can assume that Ω is bounded. Let us consider the function

$$G(x, s, \xi) \equiv F(x, s + D^{[k-1]}\phi(x), D^k\phi(x) + \xi).$$

G is a continuous function, coercive in the rank-1 direction Λ and quasi-convex with respect to the variable ξ, and

$$G(x, 0, 0) = F(x, D^{[k-1]}\phi(x), D^k\phi(x)) < 0 \quad \forall x \in \overline{\Omega}.$$

So our problem is equivalent to the following one: find

$$\begin{cases} w \in W_0^{k,\infty}, \\ G(x, D^{[k-1]}w(x), D^k w(x)) = 0, \quad \text{a.e. } x \in \Omega. \end{cases} \tag{5.2}$$

Let $r > 0$. If $v \in W_0^{k,\infty}(\Omega)$ and $|D^k v(x)| \le r$, a.e. $x \in \Omega$, we know that $\|D^{[k-1]}v\|_\infty \le Lr$, $L = L(\mathrm{diam}(\Omega))$. We consider the coercivity condition (2.3) with $x \in \overline{\Omega}$, $|s| \le Lr$, $|\xi| \le r$:

There exist m, $q \in \mathbb{R}$, $m > 0$ such that $G(x, s, \xi + t\Lambda) \ge m|t| - q \ \forall t \in \mathbb{R}$. We define

$$K = \left\{\eta \in ((\mathbb{R}^n)^{\otimes_k})_s \colon \eta = \xi + t\Lambda, \quad |\xi| \le r, \quad t \in \mathbb{R} \colon m|t| - q \le 0\right\}.$$

K is a convex compact set and

$$\left\{\eta \in ((\mathbb{R}^n)^{\otimes_k})_s \colon \eta = \xi + t\Lambda, \quad |\xi| \le r, \quad G(x, s, \eta) \le 0 \quad \forall x \in \overline{\Omega} \quad \forall s \colon |s| \le Lr\right\} \subset K.$$

Let

$$W = \Big\{u \in W_0^{k,\infty}(\Omega) \colon \exists \varepsilon_l \downarrow 0 \exists u_l \in P_{\varepsilon_l}(\Omega) \text{ such that}$$

$$\begin{cases} \|u - u_l\|_{k-1,\infty,\Omega} \le \varepsilon_l, \\ D^k u_l(x) \in \mathrm{int}K, & \text{a.e. } x \in \Omega, \\ G(x, D^{[k-1]}u_l(x), D^k u_l(x)) < 0 & \forall x \in \Omega, \\ u_l \in W_0^{k,\infty}(\Omega). \end{cases}$$

Since $\psi \equiv 0$ is in W, W is nonempty and, as in Lemma 4.1, $(W, C^{k-1}(\overline{\Omega}))$ is a complete metric space. Moreover, since G is quasi-convex we have

$$W \subset \{u \in W_0^{k,\infty}(\Omega) \colon G(x, D^{[k-1]}u(x), D^k u(x)) \le 0, \text{ a.e. } x \in \Omega\};$$

see [18], [13], [15]. For each $m \in \mathbb{N}$ let us consider the set

$$W^m = \left\{u \in W \colon \int_\Omega G(x, D^{[k-1]}u(x), D^k u(x)) \ge \frac{-1}{m}\right\}.$$

W^m is an open subset of W; indeed, since G is quasi-convex, $W \setminus W^m$ is closed.

We want to prove that W^m is also a dense subset of $(W, C^{k-1}(\Omega))$. Let $w \in W$; by definition there exist $\varepsilon > 0$ and w_ε such that

$$\begin{aligned} &w_\varepsilon \in P_\varepsilon(\Omega), \\ &D^k w_\varepsilon(x) \in \mathrm{int}K, \ \text{ a.e. } x \in \Omega, \\ &G(x, D^{[k-1]}w_\varepsilon(x), D^k w_\varepsilon(x)) < 0, \ \text{ a.e. } x \in \Omega, \\ &w_\varepsilon \in \phi + W_0^{k,\infty}(\Omega), \\ &\|w_\varepsilon - w\|_{k-1,\infty,\Omega} < \frac{\varepsilon}{2}. \end{aligned}$$

By definition of $P_\varepsilon(\Omega)$, there exist Ω_0, Ω_j, $j \in \mathbb{N}$ open Lipschitz pairwise disjoint subsets of Ω such that

$$\begin{aligned} &\bigcup_{j \ge 0} \overline{\Omega_j} = \overline{\Omega}, \\ &|\Omega_0| \le \varepsilon, \\ &D^k w_\varepsilon(x) = \xi_j \in \ \mathrm{int}K, \ \text{ a.e. } x \in \Omega_j \quad \forall j \in \mathbb{N}, \\ &\xi_j \in \ \mathrm{int}K, \\ &G(x, D^{[k-1]}w_\varepsilon(x), \xi_j) < 0 \quad \forall x \in \Omega_j \quad \forall j \in N. \end{aligned}$$

Since G is continuous and $\overline{\Omega_j}$ is compact, there exists $\delta_j > 0$ such that $\delta_j \le \delta\ \forall j \in \mathbb{N}$ such that

$$G(x, D^{[k-1]}w_\varepsilon(x), \xi_j) < -\delta_j \quad \forall x \in \overline{\Omega_j}.$$

The functions $D^{[k-1]}v(\bullet)$ are equicontinuous for $v \in W$; hence there exist Ω_{jh}, $h = 1, \dots, H_j$, open Lipschitz pairwise disjoint subsets such that $\bigcup_{h=1}^{H_j} \overline{\Omega_{jh}} = \overline{\Omega_j}$ and

$$\begin{aligned}(5.3)\quad & |G(x_1, D^{[k-1]}v_1(x_1), \xi) - G(x_2, D^{[k-1]}v_2(x_2), \xi)| < -\delta_j \qquad \forall x_1, x_2 \in \Omega_{jh},\\ & \forall v_1, v_2 \in W^{k,\infty}(\Omega_{jh}), \quad v_1 - v_2 \in W_0^{k,\infty}(\Omega_{jh}),\\ & D^k v_1(x), D^k v_2(x) \in K,\\ & \text{a.e. } x \in \Omega_{jh}, \quad \text{and} \quad \forall \xi \in K.\end{aligned}$$

For each $h = 1, \dots, H_j$ let us fix $x_h \in \Omega_{jh}$; we have

$$(5.4)\qquad G(x_h, D^{[k-1]}w_\varepsilon(x_h), \xi_j) < -\delta_j \quad \forall h = 1, \dots, H_j.$$

We can therefore solve the following problem:

$$(5.5)\qquad \begin{cases} G(x_h, D^{[k-1]}w_\varepsilon(x_h), D^k v(x)) = -\delta_j, & \text{a.e. } x \in \Omega_{jh},\\ v \in w_\varepsilon + W_0^{k,\infty}(\Omega_{jh}). \end{cases}$$

By Lemma 4.1, in fact there exists a solution v_{jh} to problem (5.5) and v_{jh} has a sequence of approximating functions. More precisely, there exists $v_{jhl} \in P_{\varepsilon_l}(\Omega_{jh})$ such that

$$\begin{cases} \|v_{jhl} - v_{jh}\|_{k-1,\infty,\Omega} \le \varepsilon_l,\\ D^k v_{jhl}(x) \in \operatorname{int} K, & \text{a.e. } x \in \Omega,\\ v_{jhl} \in v_{jh} + W_0^{k,\infty}(\Omega_{jh}) = w_\varepsilon + W_0^{k,\infty}(\Omega_{jh}),\\ G(x_h, D^{[k-1]}w_\varepsilon(x_h), D^k v_{jhl}(x)) < -\delta_j, & \text{a.e. } x \in \Omega_{jh}. \end{cases}$$

By (5.3) we get

$$(5.6)\qquad G(x, D^{[k-1]}w_{jhl}(x), D^k v_{jhl}(x)) < -0, \text{ a.e. } x \in \Omega_{jh}.$$

We can now define the function $v \in W^m$ approximating w:

$$(5.7)\qquad v(x) = \begin{cases} w_\varepsilon(x), & x \in \overline{\Omega_0},\\ v_{jh}(x), & x \in \Omega_{jh}, \quad h = 1, \dots, H_j,\ j \in \mathbb{N}. \end{cases}$$

So, by construction, $v \in w + W_0^{k,\infty}(\Omega) = W_0^{k,\infty}(\Omega)$.

Let us compute $\int_\Omega G(x, D^{[k-1]}v(x), D^k v(x))\,dx$:

$$\begin{aligned}&\int_\Omega G(x, D^{[k-1]}v(x), D^k v(x))\,dx\\ &= \int_{\Omega_0} G(x, D^{[k-1]}w_\varepsilon(x), D^k w_\varepsilon(x))\,dx + \sum_{j\in\mathbb{N}}\sum_{h=1}^{H_j}\int_{\Omega_{jh}} G(x, D^{[k-1]}v_{jh}(x), D^k v_{jh}(x))\,dx\\ &= \int_{\Omega_0} G(x, D^{[k-1]}w_\varepsilon(x), D^k w_\varepsilon(x))\,dx + \sum_{j\in\mathbb{N}}\sum_{h=1}^{H_j}\int_{\Omega_{jh}} G(x, D^{[k-1]}w_\varepsilon(x), D^k v_{jh}(x))\,dx\\ &\quad + \sum_{j\in\mathbb{N}}\sum_{h=1}^{H_j}\int_{\Omega_{jh}} \Big[G(x, D^{[k-1]}v_{jh}(x), D^k v_{jh}(x)) - G(x, D^{[k-1]}w_\varepsilon(x), D^k v_{jh}(x))\Big]\,dx.\end{aligned}$$

We have $\int_{\Omega_0} G(x, D^{[k-1]}w_\varepsilon(x), D^k w_\varepsilon(x))\,dx \geq -C|\Omega_0|$ because $D^k w_\varepsilon(x)$ is in the compact set K a.e. $x \in \Omega_0$ and G is continuous;

$$\sum_{j\in\mathbb{N}}\sum_{h=1}^{H_j}\int_{\Omega_{jh}}\Big[G(x, D^{[k-1]}v_{jh}(x), D^k v_{jh}(x)) - G(x, D^{[k-1]}w_\varepsilon(x), D^k v_{jh}(x))\Big]\,dx$$
$$\geq \sum_{j\in\mathbb{N}} -\delta_j|\Omega_j| \geq -\delta|\Omega| \text{ by (5.3);}$$
$$\sum_{j\in\mathbb{N}}\sum_{h=1}^{H_j}\int_{\Omega_{jh}} G(x, D^{[k-1]}w_\varepsilon(x), D^k v_{jh}(x))\,dx = -\delta_j \text{ by construction.}$$

Therefore

$$\int_\Omega G(x, D^{[k-1]}v(x), D^k v(x))\,dx \geq -C\varepsilon - 2|\Omega|\delta \geq \frac{-1}{m}$$

if ε and δ are small enough. The thesis follows from Baire's lemma. □

We now shall prove our main theorem.

THEOREM 5.2 (Theorem 1.1). *Let $\Omega \subset \mathbb{R}^n$ be an open Lipschitz set. Let $F\colon \Omega \times \cdots \times ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$ be a continuous function, quasi-convex and coercive in a* rank-1 *direction Λ with respect to the last variable. Let $\phi\colon \overline{\Omega} \to \mathbb{R}$ be a $C^k(\overline{\Omega})$ function such that $F(x, D^{[k-1]}\phi(x), D^k\phi(x)) \leq 0 \quad \forall x \in \overline{\Omega}$.*

Then there exists a function $u \in \phi + W_0^{k,\infty}(\Omega)$ such that

$$F(x, D^{[k-1]}u(x), D^k u(x)) = 0, \qquad \text{a.e. } x \in \Omega. \tag{5.8}$$

Proof. As in the previous lemmas we can assume that Ω is bounded. Let us define

$$\Omega_0 \equiv \Big\{x \in \Omega\colon F(x, D^{[k-1]}\phi(x), D^k\phi(x)) = 0\Big\}.$$

Since $D^k\phi$ and F are continuous, Ω_0 is closed, hence $\Omega \setminus \Omega_0$ is open. It might not be a Lipschitz set, but $\Omega \supset \Omega_0$, hence $D^k\phi(x)$ is defined on $\partial\Omega_0$ and we can consider the following problem:

$$\begin{cases} F(x, D^{[k-1]}u(x), D^k u(x)) = 0, & \text{a.e. } x \in \Omega \setminus \Omega_0, \\ u \in \phi + W_0^{k,\infty}(\Omega_0). \end{cases} \tag{5.9}$$

The boundary datum ϕ doesn't satisfy the compatibility condition

$$F(x, D^{[k-1]}\phi(x), D^k\phi(x)) < 0 \quad \forall x \in \overline{\Omega \setminus \Omega_0}.$$

Nevertheless we can solve (5.9): For $t > 0$ let us consider the set

$$\Omega^t \equiv \Big\{x \in \Omega \setminus \Omega_0\colon F(x, D^{[k-1]}\phi(x), D^k\phi(x)) = t\Big\}.$$

We want to prove that the set

$$\big\{t < 0\colon |\Omega^t| = 0\big\}$$

is dense in $(-\infty, 0)$. Let

$$T_l \equiv \left\{t < 0\colon \frac{|\Omega \setminus \Omega_0|}{l+1} < |\Omega^t| < \frac{|\Omega \setminus \Omega_0|}{l}\right\}.$$

We have

$$\Omega \setminus \Omega_0 = \bigcup_{t<0} \Omega^t \supset \bigcup_{t\in T_l} \Omega^t$$

which implies

$$+\infty > |\Omega \setminus \Omega_0| > \sum_{t\in T_l} |\Omega^t| > \frac{\Omega \setminus \Omega_0}{l+1} \#(T_l).$$

Hence T_l must be a finite set and $\bigcup_{l\in\mathbb{N}} T_l$ is countable, therefore its complementary set $\{t<0\colon |\Omega^t| = 0\}$ is dense in $(-\infty, 0)$. In particular there exists a sequence $t_l \uparrow 0$ such that $|\Omega^{t_l}| = 0$. Let us define

$$\Omega_l \equiv \left\{ x \in \Omega \setminus \Omega_0 \colon t_l < F(x, D^{[k-1]}\phi(x), D^k\phi(x)) < t_{l+1} \right\}. \tag{5.10}$$

We can find a solution u_l to the problem

$$\begin{cases} F(x, D^{[k-1]}u(x), D^k u(x)) = 0, & \text{a.e. } x \in \Omega_l, \\ u \in \phi + W_0^{k,\infty}(\Omega_l). \end{cases} \tag{5.11}$$

Let us define

$$u(x) = \begin{cases} \phi(x), & x \in \Omega_0, \\ u_l(x), & x \in \Omega_l. \end{cases} \tag{5.12}$$

Then u is a solution to our problem. □

6. Systems of partial differential equations. We begin this section with a structure lemma whose proof can be found in [9].

PROPOSITION 6.1. *Let $E \subset ((\mathbb{R}^n)^{\otimes_k})_s$. Let us define*

$$RcoE = E$$

and, by induction,

$$\begin{aligned} R_{i+1}coE = \big\{ \xi \in ((\mathbb{R}^n)^{\otimes_k})_s \colon \xi = tA + (1-t)B, \quad t \in [0,1], \\ A, B \in \mathbb{R}_i coE \quad \operatorname{rank}\{A - B\} = 1 \big\}. \end{aligned}$$

Then

$$RcoE = \bigcup_{i\in N} R_i coE. \tag{6.1}$$

The next lemma is an extension of Lemma 3.1.

LEMMA 6.2. *Let $E \subset ((\mathbb{R}^n)^{\otimes_k})_s$ be a bounded set. Let $\Omega \subset \mathbb{R}^n$ be an open Lipschitz set. Let $\xi \in RcoE$. Let $\phi\colon \overline{\Omega} \to \mathbb{R}$ be such that $D^k\phi(x) \equiv \xi$ for every $x \in \Omega$. Then for any $\varepsilon > 0$ there exist a function $u \in \phi + W_0^{k,\infty}(\Omega)$ and an open Lipschitz set $\widetilde{\Omega} \subset \Omega$ such that*

$$u \in P_\varepsilon(\Omega), \tag{6.2}$$

$$\|u - \phi\|_{k-1,\infty,\Omega} < \varepsilon, \tag{6.3}$$

$$|\Omega \setminus \widetilde{\Omega}| < \varepsilon, \tag{6.4}$$

$$D^k u(x) \in E, \quad \text{a.e. } x \in \widetilde{\Omega}, \tag{6.5}$$

$$\operatorname{dist}(D^k u(x), RcoE) \le \varepsilon, \quad \text{a.e. } x \in \Omega. \tag{6.6}$$

Proof. The proof consists of an iteration of Lemma 3.1. Let $\xi \in \mathrm{Rco}E = \bigcup_{i\in\mathbb{N}} \mathrm{R}_i\mathrm{co}E$; then there exists $i \in \mathbb{N}$ such that $\xi \in \mathrm{R}_i\mathrm{co}E$. If $i = 0$, there is nothing to prove; we only need to take $u = \phi$. If $i = 1$, then $\xi = tA + (1-t)B$ where $A, B \in E$, $\mathrm{rank}\{A - B\} = 1$, and $t \in [0,1]$, so we only need to apply Lemma 3.1.

We shall prove Lemma 6.2 by induction on i. Let us assume $\xi \in \mathrm{R}_{i+1}\mathrm{co}E$; then $\xi = tA + (1-t)B$ where $A, B \in \mathrm{R}_i\mathrm{co}E$, $\mathrm{rank}\{A - B\} = 1$, and $t \in [0,1]$. We apply Lemma 3.1 to A and B: there exist $v \in P_\varepsilon(\Omega)$, Ω_A, $\Omega_B \subset \Omega$ open Lipschitz subsets such that

$$
\begin{aligned}
&v \in \phi + W_0^{k,\infty}(\Omega),\\
&D^k v(x) = \begin{cases} Ax \in \Omega_A,\\ Bx \in \Omega_B,\end{cases}\\
&\mathrm{dist}(D^k v(x), \mathrm{Rco}E) < \mathrm{dist}(D^k v(x), \mathrm{R}_i\mathrm{co}E) < \varepsilon, \quad \text{a.e. } x \in \Omega,\\
&\|v - \phi\|_{k-1,\infty,\Omega} < \frac{\varepsilon}{2}.
\end{aligned}
$$

Since A, $B \in \mathrm{Rco}E$, and Ω_A, Ω_B are open Lipschitz sets where v is a polynomial whose degree is less or equal than k we have

$$
\begin{aligned}
&\exists \widetilde{\Omega_A} \subset \Omega_A, \quad \exists v_A \in P_\varepsilon(\Omega_A) \text{ such that}\\
&v_A \in v + W_0^{k,\infty}(\Omega_A),\\
&D^k v_A \in E, \text{ a.e. } x \in \widetilde{\Omega_A},\\
&\mathrm{dist}(D^k v_A(x), \mathrm{Rco}E) < \varepsilon, \text{ a.e. } x \in \Omega_A,\\
&\|v - v_A\|_{k-1,\infty,\Omega_A} < \frac{\varepsilon}{2},\\
&\exists \widetilde{\Omega_B} \subset \Omega_B, \quad \exists v_B \in P_\varepsilon(\Omega_B) \text{ such that}\\
&v_B \in v + W_0^{k,\infty}(\Omega_B),\\
&D^k v_B \in E, \text{ a.e. } x \in \widetilde{\Omega_B},\\
&\mathrm{dist}(D^k v_B(x), \mathrm{Rco}E) < \varepsilon, \text{ a.e. } x \in \Omega_B,\\
&\|v - v_B\|_{k-1,\infty,\Omega_B} < \frac{\varepsilon}{2},
\end{aligned}
$$

so we just need to set

$$
\widetilde{\Omega} = \widetilde{\Omega_A} \cup \widetilde{\Omega_B};
$$

$$
u(x) = \begin{cases} v(x), & x \in \Omega \setminus (\Omega_A \cup \Omega_B),\\ v_A(x), & x \in \Omega_A,\\ v_B(x), & x \in \Omega_B. \end{cases} \qquad \square
$$

We omit the proofs of the following lemma and theorem which can be found in Dacorogna and Marcellini [10] for the case $k = 2$.

LEMMA 6.3. *Let $\Omega \subset \mathbb{R}$ be an open Lipschitz set. Let $F_i^\delta : ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$, $i = 1, \dots, N$ be quasi-convex functions, continuous with respect to the parameter $\delta \in [0, \delta_0)$, $\delta_0 > 0$ such that*

$$
\begin{aligned}
&Rco\left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s : F_i^\delta(\xi) = 0 \quad \forall i = 1, \dots, N\right\}\\
&= \left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s : F_i^\delta(\xi) \le 0 \quad \forall i = 1, \dots, N\right\} \quad \textit{bounded set of } ((\mathbb{R}^n)^{\otimes_k})_s, \qquad (6.7)
\end{aligned}
$$

$$
\begin{aligned}
&\left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s : F_i^\delta(\xi) = 0 \quad \forall i = 1, \dots, N\right\}\\
&\subset \left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s : F_i^0(\xi) < 0 \quad \forall i = 1, \dots, N\right\} \quad \forall \delta \in (0, \delta_0). \qquad (6.8)
\end{aligned}
$$

Let $\phi\colon \overline{\Omega} \to \mathbb{R}$ be a polynomial of degree less than or equal to k such that

$$F_i^0(D^k\phi(x)) \equiv F_i^0(\xi_0) < 0 \quad \forall i = 1, \dots, N. \tag{6.9}$$

Then there exists $u \in \phi + W_0^{k,\infty}(\Omega)$ such that $F_i^0(D^k u(x)) = 0$ a.e. $x \in \Omega$ $\forall i = 1, \dots, N$.

THEOREM 6.4. *Let $\Omega \subset \mathbb{R}$ be an open Lipschitz set. Let $F_i^\delta\colon \overline{\Omega} \times \cdots \times ((\mathbb{R}^n)^{\otimes_k})_s \to \mathbb{R}$, $i = 1, \dots, N$ be continuous functions, quasi-convex functions with respect to the last variable, continuous with respect to the parameter $\delta \in [0, \delta_0)$, $\delta_0 > 0$. Moreover, let us assume that for every $(x, s) \in \overline{\Omega} \times \mathbb{R} \times \cdots \times ((\mathbb{R}^n)^{\otimes_{k-1}})_s$,*

$$\begin{aligned} &Rco\left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s \colon F_i^\delta(x, s, \xi) = 0 \quad \forall i = 1, \dots, N\right\} \\ &= \left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s \colon F_i^\delta(x, s, \xi) \le 0 \quad \forall i = 1, \dots, N\right\} \quad \textit{bounded set of } ((\mathbb{R}^n)^{\otimes_k})_s; \end{aligned} \tag{6.10}$$

$$\begin{aligned} &\left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s \colon F_i^\delta(x, s, \xi) = 0 \quad \forall i = 1, \dots, N\right\} \\ &\subset \left\{\xi \in ((\mathbb{R}^n)^{\otimes_k})_s \colon F_i^0(x, s, \xi) < 0 \quad \forall i = 1, \dots, N\right\} \quad \forall \delta \in (0, \delta_0). \end{aligned} \tag{6.11}$$

Let $\phi\colon \overline{\Omega} \to \mathbb{R}$ be a (piecewise $C^k(\overline{\Omega})$) function such that

$$F_i^0(x, D^{[k-1]}\phi(x), D^k\phi(x)) < 0 \quad \forall i = 1, \dots, N. \tag{6.12}$$

Then there exists $u \in \phi + W_0^{k,\infty}(\Omega)$ such that $F_i^0(x, D^{[k-1]}u(x), D^k u(x)) = 0$ a.e. $x \in \Omega$ $\forall i = 1, \dots, N$.

REFERENCES

[1] J. M. BALL, J. C. CURRIE, AND P. J. OLVER, *Null lagrangians, weak continuity, and variational problems of arbitrary order*, J. Funct. Anal., 41 (1981), pp. 135–174.

[2] A. BRESSAN AND F. FLORES, *On total differential inclusions*, Rend. Sem. Mat. Univ. Padova, 92 (1994), pp. 9–16.

[3] A. CELLINA, *On the differential inclusion $x' \in [-1, 1]$*, Atti Accad. Naz. Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Suppl., (1980), pp. 1–6.

[4] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math Soc., 277 (1983), pp. 1–42.

[5] B. DACOROGNA, *Weak Continuity and Weak Lower Semicontinuity of Nonlinear Functionals*, Springer-Verlag, Berlin, 1982.

[6] B. DACOROGNA AND P. MARCELLINI, *Sur le problème de Cauchy-Dirichlet pour les systèmes d'équations non-linéaires du premier ordre*, C. R. Acad. Sci. Paris. Sér. I Math., 323 (1996), pp. 599–602.

[7] B. DACOROGNA AND P. MARCELLINI, *Théorem d'existence dans le cas scalaire et vectoriel pour les équations de Hamilton-Jacobi*, C. R. Acad. Sci. Paris. Sér. I Math., 322 (1996), pp. 237–240.

[8] B. DACOROGNA AND P. MARCELLINI, *General existence theorems for Hamilton-Jacobi equations in the scalar and vectorial cases*, Acta Math., 178 (1997), pp. 1–37.

[9] B. DACOROGNA AND P. MARCELLINI, *Cauchy-Dirichlet problems for first order non linear systems*, J. Funct. Anal., 152 (1998), pp. 404–446.

[10] B. DACOROGNA AND P. MARCELLINI, *Implicit second order partial differential equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), to appear.

[11] F. S. DE BLASI AND G. PIANIGIANI, *A Baire category approach to the existence of solutions of multivalued differential equations in Banach spaces*, Funkcial. Ekvac., 25 (1982), pp. 153–162.

[12] F. S. DE BLASI AND G. PIANIGIANI, *Non convex valued differential inclusions in Banach spaces*, J. Math. Anal. Appl., (1991), pp. 469–494.

[13] N. FUSCO, *Quasi convessità e semicontinuità per integrali multipli di ordine superiore*, Ricerche Mat., 29 (1980), pp. 307–323.

[14] M. GROMOV, *Partial Differential Relations*, Springer-Verlag, Berlin, 1986.

[15] M. Guidorzi and L. Poggiolini, *Lower semicontinuity for quasiconvex integrals of higher order*, Nonlinear Differential Equations Appl., 6 (1999), pp. 227–246.
[16] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, Dover, New York, 1975.
[17] N. Meyers, *An L^p-estimate for the gradient of solutions of second order elliptic divergence equations*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 17 (1963), pp. 189–206.
[18] N. Meyers, *Quasiconvexity and lower semicontinuity of multiple integrals of any order*, Trans. Amer. Math. Soc., 119 (1965), pp. 125–149.
[19] C. B. Morrey, *Quasiconvexity and the lower semicontinuity of multiple integrals*, Pacific J. Math., 2 (1952), pp. 25–53.
[20] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, Berlin, New York, 1966.
[21] S. Müller and V. Sverak, *Attainment results for the two-well problem by convex integration*, in Geometric Analysis and the Calculus of Variations, J. Jost, ed., International Press, Cambridge, MA, 1996, pp. 239–251.
[22] S. Müller and V. Sverak, *Unexpected solutions of first and second order partial differential equations*, in Proceedings of the International Congress of Mathematicians, Berlin, 1998, Doc. Math., extra vol. II (1998), pp. 691–702.

SIAM J. MATH. ANAL.
Vol. 30, No. 5, pp. 1151–1174

SPATIAL DECAY SOLUTIONS OF THE BOLTZMANN EQUATION: CONVERSE PROPERTIES OF LONG TIME LIMITING BEHAVIOR*

XUGUANG LU†

Abstract. We prove some converse properties of long time limiting behavior (along the particle paths) of a class of spatial decay solutions of the Boltzmann equation. It is shown that different initial data f_0 determine different long time limit functions $f_\infty(x,v) = \lim_{t\to\infty} f(x+tv,v,t)$, and for any given function $F(x,v)$ which belongs to a function set, there exists a solution f such that $f_\infty = F$. Existence of such spatial decay solutions are proven for the inverse power potentials with weak angular cut-off condition and for the initial data f_0 satisfying $f_0(x,v) \le C(1+|x|^2+|v|^2)^{-k}$, or $f_0(x,v) \le C(1+|x-v|^2)^{-k}$, etc. For the soft potentials, the solutions may have "locally infinite particles," i.e., $\int_{\mathbf{R}^3} f(x,v,t)dv \equiv \infty$.

Key words. Boltzmann equation, spatial decay solution, long time limit, converse property, weak angular cut-off, initial data, final value problem

AMS subject classifications. 76P05, 82C40, 35Q21

PII. S0036141098334985

1. Introduction.

In this paper we study long time limiting behavior of spatial decay solutions of the Cauchy problem for the Boltzmann equation

$$\frac{\partial}{\partial t} f + v\cdot\nabla_x f = Q(f,f) \quad \text{in } \mathbf{R}^3\times\mathbf{R}^3\times(0,\infty), \tag{B}$$

$$f|_{t=0} = f_0 \qquad \text{on} \qquad \mathbf{R}^3\times\mathbf{R}^3,$$

which describes the time evolution of the particle number density $f = f(x,v,t)$ (at time $t\in[0,\infty)$, position $x\in\mathbf{R}^3$, and velocity $v\in\mathbf{R}^3$) of a simple monoatomic gas of identical particles. $v\cdot\nabla_x$ denotes differentiation with respect to x in the direction of v, and $Q(f,f)$ is the so-called collision integral, which describes the rate of change of f due to a binary collision. Let us first recall some basic facts about (B), which are also used later. Under some assumption of angular cut-off, the collision operator Q can be written as the difference of two positive bilinear forms (i.e., the gain term and the loss term):

$$Q(f,g)(v) = Q^+(f,g)(v) - Q^-(f,g)(v), \tag{1.1}$$

where

$$Q^+(f,g)(v) = \iint_{\mathbf{R^3}\times\mathbf{S^2}} B(v-v_*,\omega) f(v') g(v'_*) d\omega dv, \tag{1.2}$$

$$Q^-(f,g)(v) = f(v)L(g)(v), \tag{1.3}$$

$$L(g)(v) = \int_{\mathbf{R}^3} A(v-v_*)g(v_*)dv_*, \quad A(z) = \int_{\mathbf{S^2}} B(z,\omega)d\omega.$$

*Received by the editors March 5, 1998; accepted for publication (in revised form) September 24, 1998; published electronically August 26, 1999. This work was supported by the National Natural Science Foundation of China.

http://www.siam.org/journals/sima/30-5/33498.html

†Department of Applied Mathematics, Tsinghua University, Beijing 100084, People's Republic of China (xglu@math.tsinghua.edu.cn).

Of course in (B), $Q(f,f)(x,v,t)$ means $Q(f(x,\cdot,t),f(x,\cdot,t))(v)$. In (1.2) and (1.3), v, v_* are the velocities of two particles before they collide, and v', v'_* are their velocities after the collision. According to the conservation laws of momentum and kinetic energy, v', v'_* and v, v_* have the relations

$$v' + v'_* = v + v_*, \quad |v'|^2 + |v'_*|^2 = |v|^2 + |v_*|^2, \tag{1.4}$$

which are equivalent to the explicit representation:

$$v' = v - \langle v - v_*, \omega\rangle\omega, \quad v'_* = v_* + \langle v - v_*, \omega\rangle\omega, \ \ \omega \in \mathbf{S}^2, \tag{1.5}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbf{R}^3$, $|x|^2 = \langle x, x\rangle$, and $\mathbf{S}^2 = \{\omega \in \mathbf{R}^3 \mid |\omega| = 1\}$. The collision kernel $B(z,\omega)$ is a nonnegative Borel function of $|z|$ and $|\langle z,\omega\rangle|$ only. For the interaction potentials of inverse power laws, $B(z,\omega)$ takes the form (see, for instance, [Ce], [T,M])

$$B(z,\omega) = b(\theta)|z|^\gamma, \qquad \theta = \arccos(|z|^{-1}|\langle z,\omega\rangle|), \quad -3 < \gamma \le 1, \tag{1.6}$$

where the nonnegative function $b(\theta)$ is at least assumed to satisfy the weak angular cut-off assumption (for defining (1.1)):

$$B_0 := \int_0^{\pi/2} b(\theta)\sin(\theta)d\theta < \infty; \tag{1.7}$$

thus $A(z) = 4\pi B_0|z|^\gamma$ is in $L^1_{loc}(\mathbf{R}^3)$. The exponent γ is related to the models of potentials of intermolecular forces, namely, the soft potentials ($-3 < \gamma < 0$), the Maxwell molecular model ($\gamma = 0$), the hard potentials ($0 < \gamma < 1$), and the hard sphere model ($\gamma = 1$, $b(\theta) = \text{const.}\cos(\theta)$). All these potentials are simultaneously contained in a general form:

$$B(z,\omega) \le b(\theta)(|z|^{\gamma_1} + |z|^{\gamma_2}), \quad -3 < \gamma_1 \le 0 \le \gamma_2 \le 1, \tag{1.8}$$

which includes the Grad cut-off condition [Gr]:

$$B(z,\omega) \le \text{const.}\cos(\theta)(|z|^{-\delta} + |z|^{1-\delta}), \quad 0 \le \delta < 1. \tag{1.9}$$

Due to the mathematical difficulties in dealing with (B), one usually considers, after integration along the particle paths, the mild form:

$$f^\sharp(x,v,t) = f_0(x,v) + \int_0^t Q(f,f)^\sharp(x,v,s)ds, \quad t \in [0,\infty), \tag{1.10}$$

where

$$f^\sharp(x,v,t) = f(x+tv,v,t), \quad Q(f,f)^\sharp(x,v,t) = Q(f,f)(x+tv,v,t), \quad \text{etc.}$$

A measurable function f is called a (global) mild solution of (B) if it is nonnegative on $[0,\infty)\times\mathbf{R}^3\times\mathbf{R}^3$ and satisfies for almost everywhere (a.e.) $(x,v) \in \mathbf{R}^3\times\mathbf{R}^3$; $Q^+(f,f)^\sharp(x,v,t)$ and $Q^-(f,f)^\sharp(x,v,t)$ are both in $L_{loc}[0,\infty)$ and (1.10) (with $f|_{t=0} = f_0$) holds $\forall\ t \in [0,\infty)$. If f is a mild solution of (B) and satisfies that $L(f)^\sharp(x,v,t) \in L^1_{loc}[0,\infty)$ for $(x,v) \in \mathbf{R}^3\times\mathbf{R}^3$ a.e., then (1.10) can also be written as the following exponential multiplier form:

$$f^\sharp(x,v,t) = f_0(x,v)\, e^{-\int_0^t L(f)^\sharp(x,v,s)ds} + \int_0^t Q^+(f,f)^\sharp(x,v,s)\, e^{-\int_s^t L(f)^\sharp(x,v,\tau)d\tau}ds. \tag{1.11}$$

Equilibrium (mild) solutions of (B) are called local Maxwellians which have the form

$$M(x,v,t) = a(x,t)\exp\{-|v-u(x,t)|^2/c(x,t)\}, \quad a(x,t) \geq 0,\ c(x,t) > 0. \tag{1.12}$$

As is well known, many results on the global existence of strong (classical), mild, and renormalized solutions of (B) have been obtained, respectively, for certain classes of initial data [I,S], [H], [B,T], [T1], [Po], [B,P,T] and for generally large L^1 initial data [D,L 1], [L1], [L2], [L3], [M,P]. In the kinetic theory of gases, after the existence of global solutions of (B) are proven, a further problem is about their long time behavior, including those along the particle paths. In the case of periodic box, i.e., the solutions $f(x,v,t)$ are periodic in each x_i with period $T_i \in (0,\infty), 1 \leq i \leq 3$, which includes the spatially homogeneous solutions, it has been proved that [A,E,P], [W] for certain classes of such solutions f, $f(x,v,t)$ always converge in $L^1(\mathbf{T}^3 \times \mathbf{R}^3)$ ($\mathbf{T}^3 = \Pi_{i=1}^3[0,T_i]$), as $t \to \infty$, to global Maxwellians $M(v) = a\exp\{-|v-u|^2/c\}$ (i.e., the coefficients in (1.12) are constants) and therefore, due to the spatial periodicity, this implies that

$$\lim_{t\to\infty} f^\sharp(x,v,t) = M(v) \quad \text{in} \quad L^1(\mathbf{T}^3 \times \mathbf{R}^3),$$

where the constant coefficients a, u, c depend only on the initial moments $\iint_{\mathbf{T}^3 \times \mathbf{R}^3} f_0(x,v)(1, v, |v|^2)dxdv$. Even for generally large L^1 initial data, a similar result (except for the uniqueness of $M(v)$) has also been obtained in [L1].

In nonperiodic cases, the conclusion on these long time behaviors may be quite different. First of all, for a class of spatial decay solutions and for generally large L^1 solutions, we must have $\lim_{t\to\infty} f(x,v,t) = 0$ in pointwise (see, for instance, [B,P,T]) and, respectively, in $L^1(\Omega \times \mathbf{R}^3)$-norm for all bounded domain $\Omega \subset \mathbf{R}^3$ [D,L 1]. Some quantitative estimates in L^1-norm on the time decay of large L^1 (or renormalized) solutions have been also established in [Pe]. On the other hand, under the Gard cut-off condition (1.9), Toscani [T1], [T2] proved that for any $p > 1/2,\ k > 3/2$, there exists a constant $C > 0$ such that if the initial datum satisfies $f_0(x,v) \leq C(1+|x|^2)^{-p}(1+|v|^2)^{-k}$ then the corresponding mild solution satisfying $f(x,v,t) \leq 2C(1+|x-tv|^2)^{-p}(1+|v|^2)^{-k}$ exists and satisfies for some function $f_\infty \geq 0$

$$\lim_{t\to\infty} \sup_{(x,v)\in\mathbf{R}^3\times\mathbf{R}^3} |f^\sharp(x,v,t) - f_\infty(x,v)| = 0$$

or equivalently, $\lim_{t\to\infty} \sup_{(x,v)\in\mathbf{R}^3\times\mathbf{R}^3} |f(x,v,t) - f_\infty(x-tv,v)| = 0$ (see also Polewczak [Po] for more strong convergence of classical solutions). Moreover, if for some constant $\varepsilon > 0$

$$f_0(x,v) \geq \varepsilon(1+|x|^2)^{-p}(1+|v|^2)^{-k} \quad \text{on} \quad \mathbf{R}^3 \times \mathbf{R}^3$$

then there exists a constant $\varepsilon' > 0$ such that (thanks to the exponential multiplier form (1.11))

$$f_\infty(x,v) \geq \varepsilon'(1+|x|^2)^{-p}(1+|v|^2)^{-k} \quad \text{on} \quad \mathbf{R}^3 \times \mathbf{R}^3.$$

This implies that $f_\infty(x-tv,v)$ cannot be any local Maxwellian because it does not decay exponentially at infinity in the velocity variable. As a consequence (for $p > 5/2,\ k > 7/2$), if T_M is the traveling Maxwellian

$$T_M(x,v) = c_1\exp(-c_2|x-\bar{x}_0|^2 - c_3|v-\bar{v}_0|^2),\ c_i > 0,$$

determined by the moments of the initial datum f_0

$$\iint_{\mathbf{R}^3\times\mathbf{R}^3} T_M(x,v)\,\varphi\,dxdv = \iint_{\mathbf{R}^3\times\mathbf{R}^3} f_0(x,v)\,\varphi\,dxdv, \quad \varphi = 1, x, v, |x|^2, |v|^2, \tag{1.13}$$

then $\inf_{t\geq 0} H(f)(t) > H(T_M)$, where H is the Boltzmann H-functional:

$$H(f)(t) = \iint_{\mathbf{R}^3\times\mathbf{R}^3} f(x,v,t)\log f(x,v,t)dxdv \quad (= H(f^\sharp)(t)).$$

Continuing this investigation, it is natural to ask which kind of initial data f_0 can imply that the long time limits $f_\infty(x,v) = \lim_{t\to\infty} f^\sharp(x,v,t)$ of the corresponding solutions satisfy that the functions $f_\infty(x-tv,v)$ are local Maxwellians? Or on the analogy of the gross determinism for long time limits in the case of periodic box, which kind of different solutions can have the same long time limit f_∞? In this paper, our main results on these questions can be roughly stated as follows: For the collision model (1.8) ($-2 < \gamma_1 \leq 0 \leq \gamma_2 \leq 1$) with the weak angular cut-off condition (1.7), and for continuous mild solutions $f, g, \ldots$ (their initial data may be different) which have upper-bounds $C\Phi$ on $\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$ where Φ is one of the following three types of functions:

$$\Phi(x,v,t) = \left(1 + \left|\frac{x-tv-x_0}{a}\right|^2 + \left|\frac{v-v_0}{b}\right|^2\right)^{-k}, \tag{1.14}$$

(1.15)

$$\Phi(x,v,t) = \left(1 + \left|\frac{x-tv-x_0}{a}\right|^2 + \left|\frac{v-v_0}{b}\right|^2\right)^{-k} e^{-\alpha|v-v_0|^\beta}, \quad \alpha > 0, \quad 0 \leq \beta \leq 2,$$

$$\Phi(x,v,t) = \left(1 + \left|\frac{x-tv-x_0}{a} - \frac{v-v_0}{b}\right|^2\right)^{-k} \tag{1.16}$$

with constants $a > 0$, $b > 0$, $(x_0, v_0) \in \mathbf{R}^3\times\mathbf{R}^3$ and with a suitably large $k > 0$ (here the third bounds (1.16) are used only for soft potentials and Maxwell model $\gamma_1 \leq 0 = \gamma_2$), we have the following:

(1) All such solutions' long time limits f_∞ exist in pointwise; and if $f_\infty(x-tv,v)$ is a local Maxwellian M, then $f \equiv M$; if $\inf_{t\geq 0} H(f)(t) = H(T_M)$, then $f \equiv T_M(x-tv,v)$, especially, $f_0 = T_M$, where the traveling Maxwellian T_M is determined by the moment condition (1.13). Therefore, if $f_0 \neq T_M$, then $\inf_{t\geq 0} H(f)(t) > H(T_M)$.

(2) If $f_\infty = g_\infty$, then $f \equiv g$, especially, $f_0 = g_0$. That is, different initial data determine different long time limits (along the particle paths).

(3) If initial data f_0 are continuous satisfying $f_0(x,v) \leq C\Phi(x,v,0)$ with a suitable constant $C > 0$, then such continuous mild solutions exist on $\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$.

(4) For any continuous function $F(x,v)$ satisfying $C_1\Phi(x,v,0) \leq F(x,v) \leq C_2\Phi(x,v,0)$ with suitable constants $C_2 > C_1 > 0$, there exists a continuous mild solution f such that $f_\infty = F$.

The results (1), (2), and (4) give some converse properties for long time limiting behavior (along the particle paths) of spatial decay or $L^1(\mathbf{R}^3\times\mathbf{R}^3)$ solutions, and show that there are some essential differences between spatial decay and nonspatial

decay (e.g., spatially periodic) solutions. Detailed statements and proofs of these results are given in section 3 for (1), (2) and section 4 for (3), (4), where under the weak angular cut-off condition (1.7) the collision kernel $B(z,\omega)$ is taking only the inverse power laws (1.6) ($-2<\gamma\le 1$) in order to simplify our notation and proofs (including those in section 2). As one will see, this treatment on the collision model does not influence the results (1)–(4) above for the collision model (1.8)($-2<\gamma_1\le 0\le\gamma_2\le 1$) since the collision operators $Q^{\pm}$ are linear and increasing with respect to kernels B. To avoid interrupting our proofs for the main results, we first prove several technical lemmas in section 2. These lemmas are also useful for further investigation of the Boltzmann equation.

2. Some lemmas. For any $z\in\mathbf{R}^3\setminus\{0\}$, let $\mathbf{S}^1(z)=\{\omega\in\mathbf{S}^2 \mid \omega\perp z\}$, and let $d^\perp\omega$ denote the Lebesgue measure on the circle $\mathbf{S}^1(z)$, i.e.,

$$\int_{\mathbf{S}^1(z)} g(\omega)d^\perp\omega := \int_0^{2\pi} g(\cos(\phi)\,\mathrm{i}+\sin(\phi)\,\mathrm{j})d\phi,\quad g\in C(\mathbf{S}^2),$$

where $\mathrm{i},\ \mathrm{j}\in\mathbf{S}^1(z)$, $\mathrm{i}\perp\mathrm{j}$. It is easily verified that the right-hand side of the integrals is independent of the choice of i, j. The following homogeneity is obvious and will be often used in this section:

$$\mathbf{S}^1(-z)=\mathbf{S}^1(z),\quad \mathbf{S}^1(\lambda z)=\mathbf{S}^1(z),\quad z\in\mathbf{R}\setminus\{0\},\quad \lambda>0. \tag{2.1}$$

LEMMA 2.1. *Let $G\in C(\mathbf{S}^2\times\mathbf{S}^2), 0\le F\in C(\mathbf{R}^3\times\mathbf{R}^3), 0\le f\in C(\mathbf{R}^3)$, and let $\rho\ge 0$ be measurable on $(0,\infty)$. Then*

$$\int_{\mathbf{S}^2}\left[\int_{\mathbf{S}^1(\sigma)} G(\sigma,\omega)d^\perp\omega\right]d\sigma=\int_{\mathbf{S}^2}\left[\int_{\mathbf{S}^1(\omega)} G(\sigma,\omega)d^\perp\sigma\right]d\omega, \tag{2.2}$$

$$\int_{\mathbf{R^3}}\rho(|z|)\left[\int_{\mathbf{S}^1(z)} F(z,|z|\omega)d^\perp\omega\right]dz=\int_{\mathbf{R^3}}\rho(|z|)\left[\int_{\mathbf{S}^1(z)} F(|z|\omega,z)d^\perp\omega\right]dz, \tag{2.3}$$

$$\int_{\mathbf{R^3}}\rho(|z|)\left[\int_{\mathbf{S}^1(z)} f(|z|\omega)d^\perp\omega\right]dz=2\pi\int_{\mathbf{R^3}}\rho(|z|)f(z)dz. \tag{2.4}$$

Proof. Choose $\delta\in C(\mathbf{R}), \delta\ge 0$ satisfying $\mathrm{supp}\delta\subset[-1,1]$, $\int_{-1}^1\delta(t)dt=1$. Let $\delta_n(t)=n\delta(nt)$. Then, by the Fubini theorem, we have

$$\int_{\mathbf{S}^2}\left[\int_{\mathbf{S}^2} G(\sigma,\omega)\delta_n(\langle\sigma,\omega\rangle)d\omega\right]d\sigma=\int_{\mathbf{S}^2}\left[\int_{\mathbf{S}^2} G(\sigma,\omega)\delta_n(\langle\sigma,\omega\rangle)d\sigma\right]d\omega.$$

On the other hand, $\forall\ \sigma\in\mathbf{S}^2$, we compute

$$\int_{\mathbf{S^2}} G(\sigma,\omega)\delta_n(\langle\sigma,\omega\rangle)d\omega=\int_0^\pi \delta_n(\cos(\theta))\sin(\theta)\int_{\mathbf{S}^1(\sigma)} G(\sigma\,,\cos(\theta)\sigma+\sin(\theta)\omega)d^\perp\omega\,d\theta$$

$$=\int_{-1}^1\delta(t)\int_{\mathbf{S}^1(\sigma)} G(\,\sigma\,,\,(t/n)\sigma+\sqrt{1-(t/n)^2}\ \,\omega\,)d^\perp\omega\,dt\to\int_{\mathbf{S}^1(\sigma)} G(\sigma,\omega)d^\perp\omega,\quad n\to\infty.$$

Similarly,

$$\lim_{n\to\infty}\int_{\mathbf{S}^2} G(\sigma,\omega)\delta_n(\langle\sigma,\omega\rangle)d\sigma=\int_{\mathbf{S}^1(\omega)} G(\sigma,\omega)d^\perp\sigma\quad\forall\ \omega\in\mathbf{S}^2.$$

Therefore (2.2) follows from the Lebesgue dominated convergence theorem. Equation (2.3) is easily proved by first using spherical coordinate transform and the Fubini–Tonelli theorem with the two integrals, and then applying (2.2) to their inner integrals of angular variables. Equation (2.4) is a special case of (2.3). ∎

LEMMA 2.2. *Let $B(z,\omega)=\bar{B}(|z|,|z|^{-1}|\langle z,\omega\rangle|)$ be a collision kernel. Then for any nonnegative function $F\in C(\mathbf{R}^3\times\mathbf{R}^3)$,*

(2.5)
$$\iint_{\mathbf{R}^3\times\mathbf{S}^2} B(v-v_*,\omega)F(v',v'_*)d\omega\,dv_*$$

$$=2\int_0^{\pi/2}\sin(\theta)\int_{\mathbf{R}^3}\bar{B}(|z|,\cos(\theta))\left[\int_{\mathbf{S}^1(z)}F(v-\cos(\theta)z,v-\sin(\theta)|z|\omega)d^{\perp}\omega\right]dz\,d\theta$$

and $\forall\ \theta\in(0,\pi/2), v\in\mathbf{R}^3$,

(2.6)
$$\int_{\mathbf{R}^3}\bar{B}(|z|,\cos(\theta))\left[\int_{\mathbf{S}^1(z)}F(v-\cos(\theta)z\,,\,v-\sin(\theta)|z|\omega)d^{\perp}\omega\right]dz$$
$$=\int_{\mathbf{R}^3}\bar{B}(|z|,\cos(\theta))\left[\int_{\mathbf{S}^1(z)}F(v-\cos(\theta)|z|\omega\,,\,v-\sin(\theta)z)d^{\perp}\omega\right]dz.$$

Furthermore, if $B(z,\omega)=b(\theta)|z|^{\gamma}$ and $f,\ g\in C(\mathbf{R}^3)$ are nonnegative, then

$$Q^+(f,g)(v)=2\int_0^{\pi/2}b(\theta)\sin(\theta)I(f,g)(\theta,v)d\theta \tag{2.7}$$

and

$$I(f,g)(\theta,v)\equiv I(g,f)(\pi/2-\theta,v),\qquad \theta\in[0,\pi/2],\quad v\in\mathbf{R}^3, \tag{2.8}$$

where

$$I(f,g)(\theta,v)=\int_{\mathbf{R}^3}|z|^{\gamma}f(v-\cos(\theta)z)\left[\int_{\mathbf{S}^1(z)}g(v-\sin(\theta)|z|\omega)d^{\perp}\omega\right]dz. \tag{2.9}$$

Proof. Let $Q^+(F)(v)$ be the left-hand side of (2.5). By (1.5) and the spherical coordinate transform we have

$$Q^+(F)(v)=\int_0^{\infty}\int_{\mathbf{S}^2}\int_{\mathbf{S}^2}r^2\bar{B}(r,|\langle\sigma,\omega\rangle|)F(v-\langle\sigma,\omega\rangle r\omega\,,\,v-r\sigma+\langle\sigma,\omega\rangle r\omega)d\omega\,d\sigma\,dr$$
$$=\int_{\mathbf{R}^3}\left[\int_{\mathbf{S}^2}\bar{B}\left(|z|\left|\left\langle\sigma,\frac{z}{|z|}\right\rangle\right|\right)F\left(v-\left\langle\sigma,\frac{z}{|z|}\right\rangle z\,,\,v-|z|\sigma+\left\langle\sigma,\frac{z}{|z|}\right\rangle z\right)d\sigma\right]dz$$
$$=\int_{\mathbf{R}^3}\int_0^{\pi}\bar{B}(|z|,|\cos(\theta)|)\sin(\theta)\left[\int_{\mathbf{S}^1(z)}F(v-\cos(\theta)z\,,\,v-\sin(\theta)|z|\omega)d^{\perp}\omega\right]d\theta\,dz$$

=the right-hand side of (2.5),

where the factor 2 in (2.5) is due to $\mathbf{S}^1(-z)=\mathbf{S}^1(z)$. Equations (2.6) and (2.8) follow from Lemma 2.1. Equation (2.7) is a special case of (2.5). ∎

The "reciprocity" relation (2.8) and the formula (2.4) are important for the following estimation of gain term (Lemma 2.3) because they allow us to estimate (2.8) alone, so that the angular function $b(\theta)$ is needed only to satisfy the weak angular cut-off (1.7). In this sense, the application of (2.7)–(2.9) and (2.4) are more convenient than those of the Carleman representation of gain term [Ca]; see also [Gu].

LEMMA 2.3. *Let $B(z,\omega)$ satisfy* (1.6), (1.7), *and let $k > (3+\gamma)/2$. Then there exist positive constants $C^{\pm} = C^{\pm}(B_0,\gamma,k) < \infty$ depending only on B_0,γ, and k such that if*

$$\Phi(v) = (a + b|v-u|^2)^{-k}, \qquad a > 0,\ b > 0,\ u \in \mathbf{R},^3$$

then

$$\frac{Q^{\pm}(\Phi,\Phi)(v)}{\Phi(v)} \le C^{\pm}\left(\frac{1}{a}\right)^{k-(3+\nu)/2}\left(\frac{1}{b}\right)^{(3+\gamma)/2}(a+b|v-u|^2)^{\mu/2}, \quad v \in \mathbf{R}^3, \tag{2.10}$$

where $\mu = \max\{\gamma,0\}$, $\nu = \min\{\gamma,0\}$.

Proof. Let $\hat{\Phi}(v) = (1+|v|^2)^{-k}$. Then using, for instance, (2.7), (2.9), and (2.1) with the change of integration variable z we have

$$\frac{Q^{\pm}(\Phi,\Phi)(v)}{\Phi(v)(a+b|v-u|^2)^{\mu/2}} = \left(\frac{1}{a}\right)^{k-(3+\nu)/2}\left(\frac{1}{b}\right)^{(3+\gamma)/2}\frac{Q^{\pm}(\hat{\Phi},\hat{\Phi})(w)}{\hat{\Phi}(w)(1+|w|^2)^{\mu/2}},$$

where $w = \sqrt{\frac{b}{a}}\,(v-u)$. Thus the estimate (2.10) is equivalent to its standard case, i.e., $a = b = 1$ and $u = 0$, so that in the following we can suppose that $\Phi(v) = (1+|v|^2)^{-k}$. Since

$$\frac{Q^{-}(\Phi,\Phi)(v)}{\Phi(v)} = L(\Phi)(v) = 4\pi B_0 \int_{\mathbf{R}^3} |z|^{\gamma}\,\Phi(v-z)dz$$

and, by Lemma 2.2,

$$Q^{+}(\Phi,\Phi)(v) \le 2B_0 \sup_{\theta\in[0,\pi/4]} I(\Phi,\Phi)(\theta,v),$$

we see that to prove the standard case of (2.10) it suffices to prove that there exists a positive constant $C = C(\gamma,k) < \infty$ such that

$$\int_{\mathbf{R}^3} |z|^{\gamma}\,\Phi(v-z)dz \le C(1+|v|^2)^{\mu/2}, \tag{2.11}$$

$$\sup_{\theta\in[0,\pi/4]} I(\Phi,\Phi)(\theta,v) \le C\Phi(v)(1+|v|^2)^{\mu/2}. \tag{2.12}$$

In the following, the same C always denotes different finite constants which depend only on γ and k. If $\gamma \ge 0$, then $\mu = \gamma$ and $|z|^{\gamma} \le (1+|v-z|^2)^{\gamma/2}(1+|v|^2)^{\gamma/2}$ which implies (2.11); if $-3 < \gamma < 0$, then $\mu = 0$ and

$$\int_{\mathbf{R}^3} |z|^{\gamma}\Phi(v-z)dz = \int_{|z|\le|v-z|} |z|^{\gamma}\Phi(v-z)dz + \int_{|z|>|v-z|} |z|^{\gamma}\Phi(v-z)dz$$
$$\le \int_{\mathbf{R}^3} |z|^{\gamma}\Phi(z)dz + \int_{\mathbf{R}^3} |v-z|^{\gamma}\Phi(v-z)dz = 8\pi\int_0^{\infty} r^{2+\gamma}(1+r^2)^{-k}dr < \infty;$$

(2.11) still holds. For any $z \in \mathbf{R}^3 \setminus \{0\}$ and any $\omega \in \mathbf{S}^1(z)$, we have, since $\omega \perp z$,

$$|z|^2 = |\cos(\theta)z - \sin(\theta)|z|\omega\ |^2 \le 2|v - \cos(\theta)z|^2 + 2|v - \sin(\theta)|z|\omega\ |^2, \tag{2.13}$$

$$|v - \cos(\theta)z|^2 + |v - \sin(\theta)|z|\omega\ |^2 = |v|^2 + |v - \cos(\theta)z - \sin(\theta)|z|\omega\ |^2 \ge |v|^2. \tag{2.14}$$

Using (2.13) we get

$$\begin{aligned}&\Phi(v - \cos(\theta)z)\Phi(v - \sin(\theta)|z|\omega)\\ &\le (1 + |v - \cos(\theta)z|^2 + |v - \sin(\theta)|z|\omega|^2)^{-k} \le \Phi\left(\frac{z}{\sqrt{2}}\right)\end{aligned}$$

which together with (2.9) implies that $I(\Phi, \Phi)$ is bounded:

$$I(\Phi, \Phi)(\theta, v) \le 2\pi \int_{\mathbf{R}^3} |z|^\gamma \Phi\left(\frac{z}{\sqrt{2}}\right) dz \le C.$$

Thus (2.12) holds for $|v| \le 1$. In the following we suppose that $|v| > 1$, and consider two cases for large and small θ in $[0, \pi/4]$.

Case 1. $\arctan(1/4) \le \theta \le \pi/4$. By (2.14) we have

$$\Phi(v - \cos(\theta)z)\Phi(v - \sin(\theta)|z|\omega) \le \Phi\left(\frac{v}{\sqrt{2}}\right)[\Phi(v - \cos(\theta)z) + \Phi(v - \sin(\theta)|z|\omega)].$$

Then, applying Lemma 2.1 and (2.4) we obtain

$$\begin{aligned}&I(\Phi, \Phi)(\theta, v) \le \Phi\left(\frac{v}{\sqrt{2}}\right)\\ &\times\left[\int_{\mathbf{R}^3} |z|^\gamma \int_{\mathbf{S}^1(z)} \Phi(v - \cos(\theta)z) d^\perp\omega dz + \int_{\mathbf{R}^3} |z|^\gamma \int_{\mathbf{S}^1(z)} \Phi(v - \sin(\theta)|z|\omega) d^\perp\omega dz\right]\\ &= 2\pi\, \Phi\left(\frac{v}{\sqrt{2}}\right)\left[\int_{\mathbf{R}^3} |z|^\gamma \Phi(v - \cos(\theta)z) dz + \int_{\mathbf{R}^3} |z|^\gamma \Phi(v - \sin(\theta)z) dz\right]\\ &= \pi\, \Phi\left(\frac{v}{\sqrt{2}}\right)\left[\left(\frac{1}{\cos(\theta)}\right)^{3+\gamma} + \left(\frac{1}{\sin(\theta)}\right)^{3+\gamma}\right] \int_{\mathbf{R}^3} |z|^\gamma \Phi(v - z) dz \le C\Phi(v)(1 + |v|^2)^{\mu/2},\end{aligned}$$

where the last inequality is due to (2.11) and also due to the fact that $\Phi(\frac{v}{\sqrt{2}}) \le C\Phi(v)$.

Case 2. $0 \le \theta \le \arctan(1/4)$. We have, by homogeneity (2.1),

$$\begin{aligned}I(\Phi, \Phi)(\theta, v) &= \left(\frac{1}{\cos(\theta)}\right)^{3+\gamma} \int_{\mathbf{R}^3} |z|^\gamma \Phi(v - z) \int_{\mathbf{S}^1(z)} \Phi(v - \tan(\theta)|z|\omega) d^\perp\omega\, dz\\ &\le C[I_1(\Phi, \Phi)(\theta, v) + I_2(\Phi, \Phi)(\theta, v)],\end{aligned}$$

where

$$\begin{aligned}I_1(\Phi, \Phi)(\theta, v) &= \int_{\tan(\theta)|z| \le \frac{1}{2}|v|} |z|^\gamma \Phi(v - z) \int_{\mathbf{S}^1(z)} \Phi(v - \tan(\theta)|z|\omega) d^\perp\omega\, dz,\\ I_2(\Phi, \Phi)(\theta, v) &= \int_{\tan(\theta)|z| > \frac{1}{2}|v|} |z|^\gamma \Phi(v - z) \int_{\mathbf{S}^1(z)} \Phi(v - \tan(\theta)|z|\omega) d^\perp\omega\, dz.\end{aligned}$$

By (2.11) we have

$$I_1(\Phi, \Phi)(\theta, v) \le 2\pi\, \Phi(\frac{v}{2}) \int_{\tan(\theta)|z| \le \frac{1}{2}|v|} |z|^\gamma \Phi(v - z) dz \le C\Phi(v)(1 + |v|^2)^{\mu/2}.$$

For $I_2(\Phi,\Phi)(\theta,v)$, since $\tan(\theta)|z| > \frac{1}{2}|v|$ implies $|v-z| \geq (\frac{1}{2\tan(\theta)} - 1)|v| \geq \frac{|v|}{4\tan(\theta)}$, it follows from (2.4) and (2.11) that

$$\begin{aligned} I_2(\Phi,\Phi)(\theta,v) &\leq \Phi\left(\frac{v}{4\tan(\theta)}\right)\int_{\mathbf{R}^3}|z|^\gamma \int_{\mathbf{S}^1(z)} \Phi(v-\tan(\theta)|z|\omega)d^\perp\omega\, dz \\ &= \Phi\left(\frac{v}{4\tan(\theta)}\right) 2\pi \left(\frac{1}{\tan(\theta)}\right)^{3+\gamma}\int_{\mathbf{R}^3}|z|^\gamma \Phi(v-z)dz \\ &\leq C\,\Phi\left(\frac{v}{4\tan(\theta)}\right)\left(\frac{1}{\tan(\theta)}\right)^{3+\gamma}(1+|v|^2)^{\mu/2} \\ &\leq C\left(\frac{1}{|v|}\right)^{2k}(\tan(\theta))^{2k-3-\gamma}(1+|v|^2)^{\mu/2} \leq C\,\Phi(v)(1+|v|^2)^{\mu/2}. \end{aligned}$$

Therefore, combining these estimates then leads to (2.12). ∎

LEMMA 2.4. *Let $h \in L^1[0,\infty)$ be positive and decreasing on $[0,\infty)$. Then $\forall$ $(x,v) \in \mathbf{R}^3 \times \mathbf{R}^3$*

$$\int_0^\infty h\left(\frac{|x+tv|}{\sqrt{1+t^2}}\right)\frac{dt}{1+t^2} \leq C\,(1+|x|^2+|v|^2)^{-1/2}, \tag{2.15}$$

where $C = \max\{\frac{\sqrt{2}\,\pi}{2}h(0), \sqrt{2}\,\pi\|h\|_{L^1[0,\infty)}\}$.

Proof. Denote by $I_{x,v}$ the left-hand side of (2.15), and let $\rho = \sqrt{|x|^2+|v|^2}$. Since $I_{x,v} \leq \frac{\pi}{2}h(0)$, (2.15) holds for $\rho \leq 1$. Suppose that $\rho > 1$. Let $\sin(\alpha) = |x|/\rho$ with $\alpha \in [0,\pi/2]$. Then

$$\begin{aligned} I_{x,v} &= \int_0^{\pi/2} h(|\cos(\theta)x + \sin(\theta)v|)d\theta \\ &\leq \int_0^{\pi/2} h(|\cos(\theta)|x| - \sin(\theta)|v|\,|)d\theta = \int_0^{\pi/2} h(\rho|\sin(\theta-\alpha)|)d\theta \\ &\leq \int_{-\pi/2}^{\pi/2} h\left(\frac{2}{\pi}\rho|\theta|\right)d\theta \leq \frac{\pi}{\rho}\|h\|_{L^1[0,\infty)} \leq \sqrt{2}\,\pi\|h\|_{L^1[0,\infty)}(1+\rho^2)^{-1/2}. \end{aligned}$$ ∎

Introducing polynomials

$$P(x,v,t) = 1 + \left|\frac{x-tv-x_0}{a}\right|^2 + \left|\frac{v-v_0}{b}\right|^2, \tag{2.16}$$

$$\widetilde{P}(x,v,t) = 1 + \left|\frac{x-tv-x_0}{a} - \frac{v-v_0}{b}\right|^2 \tag{2.17}$$

the functions (1.14)–(1.16) can then be written, respectively,

$$\Phi = P^{-k}, \quad P^{-k}e^{-\alpha|v-v_0|^\beta}, \quad \widetilde{P}^{-k}.$$

LEMMA 2.5. *Assume that $B(z,\omega)$ satisfy* (1.6), (1.7) *with $-2 < \gamma \leq 1$. Let $k > (3+\gamma)/2$, $\mu = \max\{\gamma,0\}$, $\nu = \min\{\gamma,0\}$, and let Φ be given by* (1.14) *or* (1.15)*, or given by* (1.16) *for $\gamma \leq 0$. Then*

(i)

$$C_p^\pm := \sup_{(x,v)\in\mathbf{R}^3\times\mathbf{R}^3} \frac{1}{\Phi(x,v,0)}\left[\int_0^\infty (1+t^2)^{\frac{3+\gamma}{2}(p-1)}(Q^\pm(\Phi,\Phi)^\sharp(x,v,t))^p dt\right]^{1/p} < \infty, \tag{2.18}$$

where $p \ge 1$ is arbitrary for $\gamma \le 0$, and $p = 1/\gamma$ for $0 < \gamma \le 1$.

(ii) *There exist positive constants $D^{\pm} = D^{\pm}(B_0, \gamma, k) < \infty$ such that $\forall a, b > 0$*

$$\sup_{(x,v)\in\mathbf{R}^3\times\mathbf{R}^3} \frac{1}{\Phi(x,v,0)} \int_0^\infty Q^{\pm}(\Phi,\Phi)^{\sharp}(x,v,t)dt \le D^{\pm}\, a\, b^{2+\gamma}. \tag{2.19}$$

(iii) *There exists a positive constant $C = C(B_0, \gamma, k, a, b) < \infty$ such that*

$$\sup_{(x,v)\in\mathbf{R}^3\times\mathbf{R}^3} \frac{1}{\Phi(x,v,0)} \int_t^\infty Q^{\pm}(\Phi,\Phi)^{\sharp}(x,v,s)ds \le C\,(1+t)^{-2-\nu}, \quad t \ge 0, \tag{2.20}$$

$$Q^{\pm}(\Phi,\Phi)(x,v,t) \le C\,(1+t^2)^{-(3+\gamma)/2}\,\Phi(x,v,t)(P(x,v,t))^{\mu/2}, \tag{2.21}$$

$$L(\Phi)(x,v,t) \le C\,(1+t^2)^{-(3+\nu)/2}(1+|v-v_0|^{\mu}). \tag{2.22}$$

(iv) *If $f, g \in C(\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty))$ are nonnegative and satisfy $f, g \le C\Phi$ for some constant $C < \infty$, then the functions*

(2.23)

$$Q^{+}(f,g), \quad L(f), \quad \int_0^t Q^{\pm}(f,g)^{\sharp}(x-tv,v,s)ds, \quad \int_t^\infty Q^{\pm}(f,g)^{\sharp}(x-tv,v,s)ds$$

are all continuous on $\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$.

Proof. (i)–(iii). Since $0 \le \beta \le 2$, we have, by the second equation (1.4),

$$|v'-v_0|^{\beta}+|v'_*-v_0|^{\beta} \ge (|v'-v_0|^2+|v'_*-v_0|^2)^{\beta/2} = (|v-v_0|^2+|v_*-v_0|^2)^{\beta/2} \ge |v-v_0|^{\beta}.$$

This implies that if Φ is given by (1.14) and Ψ is given by (1.15) then

$$\Psi(x,v',t)\Psi(x,v'_*,t) \le e^{-\alpha|v-v_0|^{\beta}}\,\Phi(x,v',t)\Phi(x,v'_*,t)$$

which together with $\Phi^{\sharp}(x,v,t) \equiv \Phi(x,v,0)$, $\Psi^{\sharp}(x,v,t) \equiv \Psi(x,v,0)$ imply

$$\frac{Q^{\pm}(\Psi,\Psi)^{\sharp}(x,v,t)}{\Psi(x,v,0)} \le \frac{Q^{\pm}(\Phi,\Phi)^{\sharp}(x,v,t)}{\Phi(x,v,0)}.$$

Therefore we need only to prove (2.18)–(2.22) for functions (1.14) and (1.16). Suppose first that Φ is given by (1.14), i.e., $\Phi = (P)^{-k}$, where P is defined by (2.16). Then using the identity

$$P(x,v,t) \equiv 1 + \frac{|x-tv_0-x_0|^2}{a^2+b^2\,t^2} + \left(\frac{t^2}{a^2}+\frac{1}{b^2}\right)|v-u(x,t)|^2, \tag{2.24}$$

where $u(x,t) = (a^2+b^2t^2)^{-1}(b^2t(x-x_0)+a^2v_0)$, and using Lemma 2.3 we have

$$\begin{aligned} &Q^{\pm}(\Phi,\Phi)(x,v,t)/\Phi(x,v,t) = Q^{\pm}(\Phi(x,\cdot,t),\Phi(x,\cdot,t))(v)/\Phi(x,v,t) \\ &\le C^{\pm}\left(1+\frac{|x-tv_0-x_0|^2}{a^2+b^2\,t^2}\right)^{-k+(3+\nu)/2}\left(\frac{t^2}{a^2}+\frac{1}{b^2}\right)^{-(3+\gamma)/2}(P(x,v,t))^{\mu/2}, \end{aligned} \tag{2.25}$$

where $C^{\pm} = C^{\pm}(B_0,\gamma,k)$ are the constants in Lemma 2.3. Obviously, (2.25) implies that (2.18)–(2.20) and (2.22) hold for $\gamma \le 0$ (i.e., for $\mu = 0, \nu = \gamma$), and (2.21) holds $\forall\ -2 < \gamma \le 1$. Now let $\gamma > 0$. Then $\mu = \gamma,\ \nu = 0$, and $p = 1/\gamma$. Let

$$h(r) = \left(\frac{1}{1+r^2}\right)^{\frac{k-3/2}{\gamma}}, \quad \widetilde{x} = \frac{x-x_0}{a}, \quad \widetilde{v} = \frac{v-v_0}{b}.$$

Then $h \in L^1[0,\infty)$ since $k > (3+\gamma)/2$ and so by (2.25) and Lemma 2.3 we have (with different constants $C_{\gamma,k,a,b}$)

$$\begin{aligned}
&\frac{1}{\Phi(x,v,0)}\left[\int_0^\infty (1+t^2)^{\frac{3+\gamma}{2}(\frac{1}{\gamma}-1)}(Q^{\pm}(\Phi,\Phi)^\sharp(x,v,t)\,)^{\frac{1}{\gamma}}\,dt\right]^\gamma\\
&\le C_{\gamma,k,a,b}\left[\int_0^\infty h\Big(\frac{|x-x_0+t(v-v_0)|}{\sqrt{a^2+b^2\,t^2}}\Big)\Big(\frac{t^2}{a^2}+\frac{1}{b^2}\Big)^{-(3+\gamma)/2}dt\right]^\gamma (P(x,v,0))^{\gamma/2}\\
&= C_{\gamma,k,a,b}\left[\int_0^\infty h\Big(\frac{|\widetilde{x}+t\widetilde{v}|}{\sqrt{1+t^2}}\Big)\Big(\frac{1}{1+t^2}\Big)^{(3+\gamma)/2}dt\right]^\gamma (P(x,v,0))^{\gamma/2}\\
&\le C_{\gamma,k,a,b}\left[\int_0^\infty h\Big(\frac{|\widetilde{x}+t\widetilde{v}|}{\sqrt{1+t^2}}\Big)\frac{dt}{1+t^2}\right]^\gamma (P(x,v,0))^{\gamma/2} \le C_{\gamma,k,a,b}.
\end{aligned}$$

Similarly, for $0<\gamma<1$ we have by the Hölder inequality $\forall\ t\ge 0$

$$\begin{aligned}
&\frac{1}{\Phi(x,v,0)}\int_t^\infty Q^{\pm}(\Phi,\Phi)^\sharp(x,v,s)ds\\
&\le C^{\pm}\,a\,b^{2+\gamma}\int_{\frac{b}{a}t}^\infty \left(h\Big(\frac{|\widetilde{x}+s\widetilde{v}|}{\sqrt{1+s^2}}\Big)\right)^\gamma\Big(\frac{1}{1+s^2}\Big)^{(3+\gamma)/2}ds\,(P(x,v,0))^{\gamma/2}\\
&\le C^{\pm}\,a\,b^{2+\gamma}\left[\int_{\frac{b}{a}t}^\infty h\Big(\frac{|\widetilde{x}+s\widetilde{v}|}{\sqrt{1+s^2}}\Big)\frac{ds}{1+s^2}\right]^\gamma\left[\int_{\frac{b}{a}t}^\infty\Big(\frac{1}{1+s^2}\Big)^{\frac{3-\gamma}{2(1-\gamma)}}ds\right]^{1-\gamma}(P(x,v,0))^{\gamma/2}\\
&\le C^{\pm}\,a\,b^{2+\gamma}\,C_{\gamma,k}\left[\int_{\frac{b}{a}t}^\infty\Big(\frac{1}{1+s^2}\Big)^{\frac{3-\gamma}{2(1-\gamma)}}ds\right]^{1-\gamma}\le C^{\pm}_{\gamma,k}a\,b^{2+\gamma}\,\Big(1+\frac{b}{a}t\Big)^{-2},
\end{aligned}$$

and for $\gamma=1$ using inequality $(1+s^2)^{-2}\le(1+s^2)^{-1}(1+(\frac{b}{a}t)^2)^{-1}\quad(s\ge\frac{b}{a}t)$ we still have

$$\frac{1}{\Phi(x,v,0)}\int_t^\infty Q^{\pm}(\Phi,\Phi)^\sharp(x,v,s)ds\le C^{\pm}_{1,k}a\,b^3\,\left(1+\Big(\frac{b}{a}t\Big)^2\right)^{-1},t\ge 0.$$

Next, by inequality $A(v-v_*)=4\pi B_0|v-v_*|^\gamma\le C\,(1+|v-v_0|^\gamma)(P(t,x,v_*))^{\gamma/2}$ and by identity (2.24) we have

$$\begin{aligned}
L(\Phi)(x,v,t)&\le C\,(1+|v-v_0|^\gamma)\Big(\frac{t^2}{a^2}+\frac{1}{b^2}\Big)^{-3/2}\int_{\mathbf{R}^3}(1+|z|^2)^{-k+\gamma/2}dz\\
&\le C(1+t^2)^{-3/2}(1+|v-v_0|^\gamma).
\end{aligned}$$

Therefore (2.18)–(2.20) and (2.22) also hold for $\gamma>0$. Now suppose that Φ is given by (1.16), i.e., $\Phi=\widetilde{P}^{-k}$, where $\widetilde{P}$ is defined by (2.17). This case corresponds to $\gamma\le 0$ by our assumption. Then starting from the identity

$$\widetilde{P}(x,v,t)\equiv 1+\Big(\frac{t}{a}+\frac{1}{b}\Big)^2|v-u(x,t)|^2,$$

where $u(x,t)=(a+bt)^{-1}(b(x-x_0)+av_0)$, and using only Lemma 2.3 we easily obtain (2.18)–(2.22).

(iv) Let $(x,v,t),(x_n,v_n,t_n)\in \mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$ and $(x_n,v_n,t_n)\to(x,v,t)$ $(n\to\infty)$. By Lemma 2.2 we have

$$Q^+(f,g)(x_n,v_n,t_n)=2\int_0^{\pi/2}b(\theta)\sin(\theta)\int_{\mathbf{R}^3}|z|^\gamma F_n(\theta,z)dz\,d\theta,$$

where $F_n(\theta,z)=F(\theta,z,x_n,v_n,t_n)$,

$$F(\theta,z,x,v,t)=f(x,v-\cos(\theta)z,t)\int_{\mathbf{S}^1(z)}g(x,v-\sin(\theta)|z|\omega,t)d^\perp\omega,\quad z\neq 0.$$

The continuity of f,g implies that $\lim_{n\to\infty}F_n(\theta,z)=F(\theta,z,x,v,t)\ \forall\ \theta\in[0,\pi/2],z\in\mathbf{R}^3\setminus\{0\}$. We may assume that $|x_n-x|+|v_n-v|\le 1$. Then by definition of Φ we have for some $0<C_{x,v}<\infty,\Phi(x_n,v_n-z,t_n)\le C_{x,v}\hat\Phi(z)$, where $\hat\Phi(z)=(1+|z|^2)^{-k}$. Since $0\le f,g\le C\Phi$, this gives

$$F_n(\theta,z)\le C^2C_{x,v}^2\,\hat\Phi(\cos(\theta)z)\int_{\mathbf{S}^1(z)}\hat\Phi(\sin(\theta)|z|\omega)d^\perp\omega.$$

Since the integral $Q^+(\hat\Phi,\hat\Phi)(0)<\infty$, it follows from the dominated convergence theorem that $\lim_{n\to\infty}Q^+(f,g)(x_n,v_n,t_n)=Q^+(f,g)(x,v,t)$. Thus $Q^+(f,g)$ is continuous on $\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$. The continuity of $L(f)$ is obvious. Finally, using the estimate (2.21) we have $\forall\ (x,v,t)\in\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$,

$$Q^\pm(f,g)^\sharp(x-tv,v,s)\le C(1+s^2)^{-(3+\gamma)/2},\quad s\in[0,\infty).$$

This and the continuity of $Q^\pm(f,g)$ imply the continuity of the third and fourth functions in (2.23). □

LEMMA 2.6. *Let f,f_*,f',f'_* be nonnegative and $\phi,\phi_*,\phi',\phi'_*$ be positive real numbers. Then*

$$\left[(ff_*-f'f'_*)\log\frac{(f'+\phi')(f'_*+\phi'_*)}{(f+\phi)(f_*+\phi_*)}\right]^+\le f'\phi'_*+f'_*\phi'+f\phi_*+f_*\phi+\phi'\phi'_*+\phi\phi_*,$$

$$\left[(f'f'_*-ff_*)\log\frac{(f'+\phi')(f'_*+\phi'_*)}{(f+\phi)(f_*+\phi_*)}\right]^+\le E(f'f'_*\ ,\ ff_*)+f'\phi'_*+f'_*\phi'+f\phi_*+f_*\phi,$$

where $(y)^+=\max\{y,0\}$ and $E(\cdot,\cdot)\ge 0$ is given by

$$E(a,b)=\begin{cases}(a-b)\log(\frac{a}{b}), & a>0,\ b>0;\\ \infty, & a>0,\ b=0\quad\text{or}\quad a=0,\ b>0;\\ 0, & a=b=0.\end{cases}$$

Proof. Denote by Δ_1,Δ_2 the left-hand sides of the two inequalities above, respectively. By symmetry and definition of E, consider, only the case that $(f'+\phi')(f'_*+\phi'_*)\ge(f+\phi)(f_*+\phi_*)$ and $ff_*>0$. If $f'f'_*\ge ff_*$, then $\Delta_1=0$ and, since $\log(1+y)\le y$,

$$\begin{aligned}\Delta_2&\le E(f'f'_*\ ,\ ff_*)+(f'f'_*-ff_*)[\log(1+\phi'/f')+\log(1+\phi'_*/f'_*)]\\&\le E(f'f'_*\ ,\ ff_*)+f'_*\phi'+f'\phi'_*.\end{aligned}$$

If $f'f'_*<ff_*$, then $\Delta_2=0$ and

$$\begin{aligned}\Delta_1&\le(ff_*-f'f'_*)\left[\ \frac{(f'+\phi')(f'_*+\phi'_*)}{(f+\phi)(f_*+\phi_*)}-\ 1\ \right]\\&\le(f'+\phi')(f'_*+\phi'_*)-(f+\phi)(f_*+\phi_*)\le f'\phi'_*+f'_*\phi'+\phi'\phi'_*.\end{aligned}$$

□

LEMMA 2.7 (see [Cs], [Ku]). *Let $\mathcal{D} \subset \mathbf{R}^N$ be a measurable set and let $F, \varphi : \mathcal{D} \to [0,\infty)$ be measurable functions satisfying $F(1+|\log F|) \in L^1(\mathcal{D})$ and*

$$\int_{\mathcal{D}} F(z)dz = \int_{\mathcal{D}} M(z)dz > 0, \quad \int_{\mathcal{D}} F(z)\varphi(z)dz \le \int_{\mathcal{D}} M(z)\varphi(z)dz < \infty,$$

where $M(z) = C\exp(-\varphi(z)), C$ is a positive constant. Then for the H-functional $H(f) = \int_{\mathcal{D}} f(z)\log f(z)dz$,

$$H(F) \ge H(M) + [\,2\|M\|_{L^1(\mathcal{D})}\,]^{-1}[\,\|F-M\|_{L^1(\mathcal{D})}\,]^2. \tag{2.26}$$

Equation (2.26) is the Csiszar–Kullback inequality which can be directly proved by the following elementary inequality:

$$|b-a| \le \left(\frac{4}{3}a + \frac{2}{3}b\right)^{1/2} [\,b\log b - a\log a - (1+\log a)(b-a)\,]^{1/2}, \quad a>0,\ b\ge 0.$$

The last lemma below is a Gronwall-type inequality, which in this paper is essentially used to deal with the case of $\gamma = 1$ of the collision model (1.6) for proving converse properties of the long time limits f_∞.

LEMMA 2.8. *Let κ be nonnegative and c be positive constants.*
(i) *Let $0 < T \le \infty, 0 \le \varrho \in L^1(0,T)$, and $0 \le u \in L^\infty(0,T)$ satisfy $\forall R \in [0,\infty)$*

$$u(t) \le \kappa + R\int_0^t \varrho(s)u(s)ds + ce^{-R}, \qquad t \in (0,T). \tag{2.27}$$

Then

$$u(t) \le \kappa + c(T)\kappa^{\theta(T)}, \qquad t \in (0,T), \tag{2.28}$$

where

$$\theta(T) = \frac{1}{2}\exp\left(-\int_0^T \varrho(s)ds\right), \quad c(T) = (ec)^{1-\theta(T)}\left(1+\int_0^T \varrho(s)ds\right).$$

(ii) *Let $0 \le \varrho \in L^1(0,\infty), 0 \le u \in L^\infty(0,\infty)$ satisfy $\forall R \in [0,\infty)$*

$$u(t) \le \kappa + R\int_t^\infty \varrho(s)u(s)ds + ce^{-R}, \qquad t \in (0,\infty).$$

Then

$$u(t) \le \kappa + c_\infty \kappa^\theta, \qquad t \in (0,\infty),$$

where

$$\theta = \frac{1}{2}\exp\left(-\int_0^\infty \varrho(s)ds\right), \quad c_\infty = (ec)^{1-\theta}\left(1+\int_0^\infty \varrho(s)ds\right).$$

Proof. By substitution $\tilde{u}(t) = u(\frac{1}{t}), \tilde{\varrho}(t) = \frac{1}{t^2}\varrho(\frac{1}{t}), t \in (0,\infty)$, part (ii) is reduced to part (i) with $T = \infty$. Thus we need only to prove part (i). In the following we set $\theta = \theta(T)$. Taking $R = 0$ in (2.27) we first obtain $u(t) \le \kappa + c\ \forall t \in (0,T)$. If

$\kappa \geq ce^{-\frac{1}{2\theta}}$, then $c \leq c(\frac{\kappa}{c}\, e^{\frac{1}{2\theta}})^{\theta} \leq (ec)^{1-\theta}\kappa^{\theta}$ and so $u(t) \leq \kappa + c(T)\kappa^{\theta}, t \in (0,T)$. Now we suppose that $\kappa < c\, e^{-\frac{1}{2\theta}}$. Let

$$C = ec\left(1 + \int_0^T \varrho(s)ds\right), \quad U_\delta(t) = \frac{1}{C}\int_0^t \varrho(s)u(s)ds + \delta, \quad t \in [0,T),$$

where δ is a positive constant. By definition of $\theta = \theta(T)$ we have $\int_0^T \varrho(s)ds < e^{\frac{1}{2\theta}}$ which together with the bounds $u(t) \leq \kappa + c$ and $\kappa < c\, e^{-\frac{1}{2\theta}}$ imply that for sufficiently small $\delta > 0$

$$\delta \leq U_\delta(t) \leq 1/e,\ t \in [0,T); \qquad \delta \leq \kappa_\delta := \frac{\kappa}{C}\int_0^T \varrho(s)ds + \delta \leq e^{-\frac{1}{2\theta}}. \tag{2.29}$$

Now taking $R = -\log(eU_\delta(t))$ in (2.27) and noticing that $ce^{-R} = ceU_\delta(t) \leq CU_\delta(t)$ we obtain

$$u(t) \leq \kappa + CU_\delta(t)|\log U_\delta(t)|, \quad t \in (0,T). \tag{2.30}$$

Multiplying $\varrho(t)/C$ to both sides of (2.30) and taking integration leads to

$$U_\delta(t) \leq \kappa_\delta + \int_0^t \varrho(s)U_\delta(s)|\log U_\delta(s)|ds, \quad t \in [0,T).$$

Let

$$G(U) = \int_\delta^U \frac{dy}{y|\log y|} = -\log\left(\frac{\log U}{\log \delta}\right), \quad U \in [\delta,\ 1/e].$$

Then $G^{-1}(V) = \exp\{(\log\delta)e^{-V}\},\ V \in [0\,,\ G(1/e)]$. Since $y \mapsto y|\log y|$ is positive and increasing on $[\delta, 1/e]$, it follows from the generalized Gronwall inequality (e.g., Bihari's inequality [B,B]) that

$$U_\delta(t) \leq G^{-1}\left(G(\kappa_\delta) + \int_0^t \varrho(s)ds\right) \leq G^{-1}\left(G(\kappa_\delta) + \int_0^T \varrho(s)ds\right) = (\kappa_\delta)^{2\theta}, \quad t \in [0,\ T).$$

Here we have checked by the second equation in (2.29) that $G(\kappa_\delta) + \int_0^T \varrho(s)ds$ lies in the domain $[0,\ G(1/e)]$ of G^{-1}. Therefore, using (2.30) and the inequality $U|\log U| \leq \sqrt{U},\ U \in [0,1]$, we obtain

$$u(t) \leq \kappa + C\sqrt{U_\delta(t)} \leq \kappa + C(\kappa_\delta)^{\theta}, \quad t \in (0,T),$$

which implies (2.28) by letting $\delta \to 0^+$. □

3. Some converse properties of long time limits. Throughout this section and the next couple of sections we will always assume that the collision kernel $B(z,\omega)$ satisfies (1.6), (1.7) with $-2 < \gamma \leq 1$. Let Φ be given by (1.14) or (1.15), or given by (1.16) for $\gamma \leq 0$. Define $\Phi_0(x,v) = \Phi(x,v,0)$ and introduce function spaces:

$$\mathcal{B}(\Phi_0) = \{f \in C(\mathbf{R}^3 \times \mathbf{R}^3) \ \mid\ \|f\|_{\Phi_0} := \sup_{(x,v)\in\mathbf{R}^3\times\mathbf{R}^3} |f(x,v)|/\Phi_0(x,v) < \infty\},$$

$$\mathcal{B}(\Phi) = \{f \in C(\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)) \ \mid\ \|f\|_{\Phi} := \sup_{t\geq 0} \|f^\sharp(t)\|_{\Phi_0} < \infty\},$$

where $f^\sharp(t) = f^\sharp(\cdot,\cdot,t)$. Note that since $\Phi(x,v,t) \equiv \Phi_0(x-tv,v)$, the norm for $\mathcal{B}(\Phi)$ can also be written

$$\|f\|_\Phi = \sup_{(x,v,t)\in \mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)} |f(x,v,t)|/\Phi(x,v,t)$$

so that both $(\mathcal{B}(\Phi), \|\cdot\|_\Phi)$ and $(\mathcal{B}(\Phi_0), \|\cdot\|_{\Phi_0})$ are Banach spaces. It is obvious that if $0 < \Psi \le \Phi$ then $\mathcal{B}(\Psi) \subset \mathcal{B}(\Phi)$. Suppose $k > (3+\gamma)/2$ and let $f \in \mathcal{B}(\Phi)$ be a mild solution of (B). Then Lemma 2.5 ensures that f satisfies (1.10) on whole $\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$ and the functions $\int_0^\infty Q^\pm(f,f)^\sharp(x,v,s)ds$, and therefore the function

$$f_\infty(x,v) := f_0(x,v) + \int_0^\infty Q(f,f)^\sharp(x,v,s)ds \tag{3.1}$$

are all in $\mathcal{B}(\Phi_0)$, where $f_0 = f|_{t=0}$. From (3.1), $f^\sharp$ can be written

$$f^\sharp(x,v,t) = f_\infty(x,v) - \int_t^\infty Q(f,f)^\sharp(x,v,s)ds. \tag{3.2}$$

This implies that f_∞ is the long time limit of $f^\sharp(t)$ in $\mathcal{B}(\Phi_0)$. In fact by Lemma 2.5 we have for a positive constant $C = C(B_0,\gamma,k,a,b) < \infty$ and for $\nu = \min\{\gamma,0\}$,

$$\|f^\sharp(t) - f_\infty\|_{\Phi_0} \le C\|f\|_\Phi^2(1+t)^{-(2+\nu)} \to 0 \qquad (t\to\infty). \tag{3.3}$$

THEOREM 3.1. *Assume that* $-2 < \gamma < 1$, $k > (3+\gamma)/2$. *Let* Φ *be given by* (1.14) *or* (1.15), *or given by* (1.16) *for* $\gamma \le 0$, *and let* $f, g \in \mathcal{B}(\Phi)$ *be mild solutions of* (B) *(their initial data may be different). Then,*

(i) *There exists* $0 < C < \infty$ *such that*

$$C^{-1}\|f_\infty - g_\infty\|_{\Phi_0} \le \|f^\sharp(t) - g^\sharp(t)\|_{\Phi_0} \le C\|f_\infty - g_\infty\|_{\Phi_0} \quad \forall\, t \ge 0. \tag{3.4}$$

(ii) *If* $f_\infty(x-tv,v)$ *is a local Maxwellian* $M(x,v,t)$, *then* $f(x,v,t) \equiv M(x,v,t)$ *on* $\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$. *In particular,* $f_0 = M_0$.

Proof. (i) From (3.2) and $ff_* - gg_* = \frac{1}{2}[(f+g)(f_*-g_*) + (f_*+g_*)(f-g)]$ we have

$$\begin{aligned}
&|f^\sharp(x,v,t) - g^\sharp(x,v,t)| \\
&\le |f_\infty(x,v) - g_\infty(x,v)| + \|f+g\|_\Phi \int_t^\infty u(s)[Q^+(\Phi,\Phi) + Q^-(\Phi,\Phi)]^\sharp(x,v,s)ds,
\end{aligned}$$

where $u(t) = \|f^\sharp(t) - g^\sharp(t)\|_{\Phi_0}$. Choose $p = 2$ for $\gamma \le 0$ and $p = 1/\gamma$ for $0 < \gamma < 1$. Let $q = p/(p-1)$, $\eta = (3+\gamma)/2$. Then by Lemma 2.5 we have

$$\begin{aligned}
&\int_t^\infty u(s)[Q^+(\Phi,\Phi) + Q^-(\Phi,\Phi)]^\sharp(x,v,s)ds \le \left[\int_t^\infty [u(s)]^q(1+s^2)^{-\eta}ds\right]^{1/q} \\
&\times\left[\int_t^\infty (1+s^2)^{\eta(p-1)}[(Q^+(\Phi,\Phi)^\sharp + Q^-(\Phi,\Phi)^\sharp)(x,v,s)]^p ds\right]^{1/p} \\
&\le \left[\int_t^\infty [u(s)]^q(1+s^2)^{-\eta}ds\right]^{1/q} (C_p^+ + C_p^-)\Phi_0(x,v).
\end{aligned}$$

Therefore, for $c = 2^{q-1}[(C_p^+ + C_p^-)\|f+g\|_\Phi]^q$,

$$[u(t)]^q \le 2^{q-1}[\|f_\infty - g_\infty\|_{\Phi_0}]^q + c\int_t^\infty [u(s)]^q(1+s^2)^{-\eta}ds, \quad t \ge 0$$

and so, by the Gronwall inequality,

$$[u(t)]^q \le 2^{q-1}[\|f_\infty - g_\infty\|_{\Phi_0}]^q \exp\{c\int_t^\infty (1+s^2)^{-\eta} ds\}, \quad t \ge 0.$$

This gives the right-hand-side inequality in (3.4) with $C = 2^{1/p}\exp\{\frac{c}{q}\int_0^\infty (1+s^2)^{-\eta} ds\}$. Similarly for each $t \ge 0$ starting from

$$f^\sharp(x,v,\tau) = f^\sharp(x,v,t) + \int_t^\tau Q(f,f)^\sharp(x,v,s) ds, \quad \tau \ge t \tag{3.5}$$

we obtain $u(\tau) \le C\|f^\sharp(t) - g^\sharp(t)\|_{\Phi_0}\ \forall\ \tau \ge t$. Applying (3.3) leads to $\lim_{\tau\to\infty} u(\tau)$ $= \|f_\infty - g_\infty\|_{\Phi_0}$ and the left-hand-side inequality in (3.4) also holds.

(ii) Suppose that $M(x,v,t) := f_\infty(x - tv, v)$ is a local Maxwellian. Then $M^\sharp(x,v,t)$ $\equiv f_\infty(x,v)$ and $M \in \mathcal{B}(\Phi)$ since $f_\infty \in \mathcal{B}(\Phi_0)$. Therefore $M_\infty = f_\infty$ and so by (3.4) $f \equiv M$ on $\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$. □

The restriction $\gamma < 1$(which excludes only the hard sphere model) will be removed in our next theorem, but the estimates like (3.4) will be given in L^1-norm for solutions $f, g \in \mathcal{B}(\Phi)$, where Φ is of type (1.15).

When dealing with the long time behavior of solutions of (B), one naturally considers the (formal) entropy equality (write $f' = f(x,v',s), f'_* = f(x,v'_*,s), f_* = f(x,v_*,s)$)

$$H(f)(t) = H(f_0) - \frac{1}{4}\int_0^t ds \iint_{\mathbf{R}^9\times\mathbf{S}^2} B(v-v_*,\omega) E(f'f'_*, ff_*) d\omega\, dv_*\, dx\, dv, t \ge 0 \tag{3.6}$$

or entropy inequality [D,L 2] (i.e., $H(f)(t) \le$ the right-hand side of (3.6)), where $E(\cdot,\cdot)$ is defined in Lemma 2.7. Unlike the cases of spatially homogeneous or spatially periodic solutions, our next theorem further shows that even though the entropy equality (3.6) can be rigorously proven for spatial decay solutions (in $\mathcal{B}(\Phi)$), this equality does not essentially help to determine what are trends of the solutions (along the particle paths).

THEOREM 3.2. *Assume that* $-2 < \gamma \le 1$. *Let* $\mu = \max\{\gamma, 0\}$.

(i) *Let* $f \in \mathcal{B}(\Phi)$ *be a mild solution of* (B), *where* Φ *is given by* (1.14) *with* $k > 3 + \frac{1}{2}\mu$. *Then the entropy equality* (3.6) *holds on* $[0,\infty)$.

(ii) *Let* Φ *be given by* (1.15) *with* $k > 3$, $\beta = \mu$, *and let* $f, g \in \mathcal{B}(\Phi)$ *be mild solutions of* (B) *(their initial data may be different). Then there exists* $0 < \theta < 1$ *and* $0 < C < \infty$ *such that*

$$[C^{-1}\|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)}]^{1/\theta} \le \|f^\sharp(t) - g^\sharp(t)\|_{L^1(\mathbf{R}^6)} \le C[\,\|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)}]^\theta \quad \forall\, t \ge 0. \tag{3.7}$$

As a consequence, the conclusion in part (ii) *of Theorem* 3.1 *still holds. Moreover, if* $k > 4$ *and if* T_M *is the traveling Maxwellian determined by* f_0 *through* (1.13), *then there exists* $0 < C < \infty$ *such that*

$$\inf_{t\ge 0} H(f)(t) = H(f_\infty) \ge H(T_M) + C\sup_{t\ge 0}[\,\|f^\sharp(t) - T_M\|_{L^1(\mathbf{R}^6)}]^{2/\theta}. \tag{3.8}$$

Proof. (i) Consider $f + \phi^n$ where $\phi^n = \frac{1}{n}\Phi$, $n \ge 1$. By $f \in \mathcal{B}(\Phi)$ and $\Phi = P^{-k}$ where P is the polynomial (2.16), one easily obtains the following estimates on $\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$:

$$(f+\phi^n)|\log(f+\phi^n)| \le C\Phi(1+\log P), \quad |\log(f+\phi^n)| \le C\, n(1+\log P) \quad \forall n \ge 1.$$

Here and below $C \in (0,\infty)$ are independent of t, x, v, and n . Since $\phi^{n\sharp}(x,v,t) = \phi_0^n(x,v)$ are independent of t, it follows that

$$\begin{aligned}&[(f+\phi^n)\log(f+\phi^n)]^\sharp(x,v,t)\\ &= [(f_0+\phi_0^n)\log(f_0+\phi_0^n)](x,v) + \int_0^t [Q(f,f)(1+\log(f+\phi^n)]^\sharp(x,v,s)ds.\end{aligned}$$

On the other hand, by Lemma 2.5 and (2.21) we have

$$Q^\pm(f,f)(x,v,t) \le \|f\|_\Phi^2 Q^\pm(\Phi,\Phi)(x,v,t) \le C(1+t^2)^{-(3+\gamma)/2}\Phi(x,v,t)[P(x,v,t)]^{\mu/2}.$$

This implies that $Q^\pm(f,f)(1+|\log(f+\phi^n)|) \in L^1(\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty))$ since $k > 3+\mu/2$. Thus according to classical derivation we have

$$H(f+\phi^n)(t) = H(f_0+\phi_0^n) - \frac{1}{4}\int_0^t ds \iint_{\mathbf{R}^9\times\mathbf{S}^2} B(v-v_*,\omega)E_n(f',f'_*,f,f_*)d\omega\, dv_*\, dx\, dv,$$

where

$$E_n(f',f'_*,f,f_*) = (f'f'_* - ff_*)\log\frac{(f'+\phi^{n\prime})(f'_*+\phi^{n\prime}{}_*)}{(f+\phi^n)(f_*+\phi^n{}_*)}.$$

Let

$$\begin{aligned}e_n^+(t) &= \frac{1}{4}\int_0^t ds \iint_{\mathbf{R}^9\times\mathbf{S}^2} B(v-v_*,\omega)[E_n(f',f'_*,f,f_*)]^+ d\omega\, dv_*\, dx\, dv,\\ e_n^-(t) &= \frac{1}{4}\int_0^t ds \iint_{\mathbf{R}^9\times\mathbf{S}^2} B(v-v_*,\omega)[-E_n(f',f'_*,f,f_*)]^+ d\omega\, dv_*\, dx\, dv.\end{aligned}$$

Then

$$e_n^+(t) = H(f_0+\phi_0^n) - H(f+\phi^n)(t) + e_n^-(t). \tag{3.9}$$

By definition of $E(\cdot,\cdot)$ it is easily shown that $\lim_{n\to\infty}[E_n(f',f'_*,f,f_*)]^+ = E(f'f'_*\,,\, ff_*)$ in pointwise. Moreover, by Lemma 2.6, we have

$$\begin{aligned}&[E_n(f',f'_*,f,f_*)]^+ \le E(f'f'_*\,,\, ff_*) + C(\Phi'\Phi'_* + \Phi\Phi_*),\\ &[-E_n(f',f'_*,f,f_*)]^+ \le \frac{C}{n}(\Phi'\Phi'_* + \Phi\Phi_*)\end{aligned} \tag{3.10}$$

so that

$$e_n^-(t) \le \frac{C}{n}\int_0^t ds \iint_{\mathbf{R}^3\times\mathbf{R}^3} Q^-(\Phi,\Phi)(x,v,s)dx\, dv \to 0 \qquad (n\to\infty).$$

Thus, by (3.9), Fatou's lemma, and the dominated convergence theorem we obtain $\forall\ t \ge 0$

$$\frac{1}{4}\int_0^t ds \iint_{\mathbf{R}^9\times\mathbf{S}^2} B(v-v_*,\omega)E(f'f'_*\,,\, ff_*)d\omega\, dv_*\, dx\, dv \le H(f_0) - H(f)(t) < \infty.$$

This integrability together with (3.10) and the dominated convergence theorem for (3.9) then implies the entropy equality (3.6).

(ii) The proof for (3.7) relies on Lemma 2.8. Let

$$u(t) = \|f^\sharp(t) - g^\sharp(t)\|_{L^1(\mathbf{R}^6)} \ \ (= \|f(\cdot,\cdot,t) - g(\cdot,\cdot,t)\|_{L^1(\mathbf{R}^6)}).$$

From (3.2) and the basic fact (see (1.4), (1.5)) that $(v, v_*) \mapsto (v', v'_*)$ is an orthogonal transform (for each fixed $\omega \in \mathbf{S}^2$) and $|v' - v'_*| = |v - v_*|, |\langle v' - v'_*, \omega\rangle| = |\langle v - v_*, \omega\rangle|$ and $B(v - v_*, \omega)$ depends only on $|v - v_*|$ and $|\langle v - v_*, \omega\rangle|$, we have

$$\begin{aligned} u(t) &\le \|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)} + \int_t^\infty ds \iint_{\mathbf{R}^3\times\mathbf{R}^3} |Q(f,f) - Q(g,g)|dx\,dv \\ &\le \|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)} + 2\int_t^\infty ds \iint_{\mathbf{R}^3\times\mathbf{R}^3} |f-g|L(f+g)dx\,dv. \end{aligned}$$

Therefore by $L(f+g) \le \|f+g\|_\Phi L(\Phi)$, Lemma 2.5, and (2.22) we obtain with $\nu = \min\{\gamma, 0\}$

$$\begin{aligned} (3.11)\ u(t) &\le \|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)} \\ &\quad + C_1 \int_t^\infty (1+s^2)^{-(3+\nu)/2} \iint_{\mathbf{R}^3\times\mathbf{R}^3} |f-g|(1+|v-v_0|^\mu)dx\,dv\,ds. \end{aligned}$$

Here and below $C_1, C_2, \ldots$ are positive and finite constants depending only on the constants $B_0, \gamma, k, a, b, \alpha$ and on the norm $\|f+g\|_\Phi$. Using the inequality

$$1 + |v - v_0|^\mu \le \left(1 + \frac{1}{\alpha}\right) R + \left(1 + \frac{1}{\alpha}\right) e^{\alpha|v-v_0|^\mu - R} \quad \forall R \ge 0$$

and recalling that $\Phi(x,v,s) = (P(x,v,s))^{-k} e^{-\alpha|v-v_0|^\mu}$ which implies

$$\begin{aligned} &\int_0^\infty (1+s^2)^{-(3+\nu)/2} \iint_{\mathbf{R}^3\times\mathbf{R}^3} |f-g| e^{\alpha|v-v_0|^\mu} dx\,dv\,ds \\ &\le \|f-g\|_\Phi \int_0^\infty (1+s^2)^{-(3+\nu)/2} ds \iint_{\mathbf{R}^3\times\mathbf{R}^3} (P(x,v,0))^{-k} dx\,dv < \infty \end{aligned}$$

we obtain by (3.11) $\forall\ R \ge 0$

$$u(t) \le \|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)} + R\int_t^\infty \varrho(s)u(s)ds + C_3\, e^{-R}, \quad t \in [0,\infty),$$

where $\varrho(s) = C_2(1+s^2)^{-(3+\nu)/2}$ $(\in L^1[0,\infty))$. It is easily seen that u is bounded and continuous on $[0,\infty)$. Then Lemma 2.8 ensures that with the number $\theta = \frac{1}{2}\exp\{-\int_0^\infty \varrho(s)ds\}$,

$$u(t) \le C[\,\|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)}\,]^\theta, \quad t \in [0,\infty).$$

In the same way, starting from (3.5) we also obtain $\forall t \ge 0, \forall R \ge 0$,

$$u(\tau + t) \le u(t) + R\int_0^\tau \varrho(s)u(s+t)ds + C_3\, e^{-R}, \quad \tau \in [0,\infty)$$

since $\varrho(s+t) \le \varrho(s)$. Therefore by Lemma 2.8 we have (with the same θ) $u(\tau+t) \le C[u(t)]^\theta\ \forall\, \tau \ge 0$. Letting $\tau \to \infty$ then leads to (by Fatou's lemma)

$$\|f_\infty - g_\infty\|_{L^1(\mathbf{R}^6)} \le C[u(t)]^\theta, \quad t \in [0,\infty).$$

These two estimates of $u(t)$ give (3.7). Finally we prove (3.8). We may suppose that $\|f_0\|_{L^1(\mathbf{R}^6)} > 0$. For otherwise, $f_0(x,v) \equiv 0$ and so applying (3.7) (take $g \equiv 0$ and then choose $t = 0,\dots$) we get $f \equiv 0$ and (3.8) is trivial. Now let $\varphi(x,v) = c_2|x - \bar{x}_0|^2 + c_3|v - \bar{v}_0|^2$ be such that $T_M(x,v) = c_1 \exp(-\varphi(x,v))$ is the traveling Maxwellian determined by f_0 through the moment condition (1.13). Since $f \in \mathcal{B}(\Phi)$, it is easily shown from Lemma 2.5 and (2.21) that $Q^{\pm}(f,f)^{\sharp}(x,v,t)(1+|x|^2+|v|^2) \in L^1(\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty))$. Thus the classical derivation shows that all moments in (1.13) are conserved by $f^{\sharp}(\cdot,\cdot,t)$. Therefore using dominated convergence we have

$$\iint_{\mathbf{R}^3\times\mathbf{R}^3} f_\infty(x,v)\{1,\varphi(x,v)\}dx\,dv = \iint_{\mathbf{R}^3\times\mathbf{R}^3} T_M(x,v)\{1,\varphi(x,v)\}dx\,dv,$$

and so by Csiszar–Kullback inequality (2.26) we obtain

$$H(f_\infty) \geq H(T_M) + [\,2\|f_0\|_{L^1(\mathbf{R}^6)}\,]^{-1}[\,\|f_\infty - T_M\|_{L^1(\mathbf{R}^6)}\,]^2. \tag{3.12}$$

Because all the coefficients c_i $(i = 1,2,3)$ are positive and $\mu \leq 1$, the solution $T_M(x - tv, v)$ is in $\mathcal{B}(\Phi)$. If we choose $g(x,v,t) = T_M(x - tv, v)$, then $g^{\sharp}(t) \equiv g_\infty = T_M$ and so (3.12) and the inequality in the right-hand side of (3.7) imply the inequality in (3.8). The equality in (3.8) is due to the monotonicity of $H(f)(t)(= H(f^{\sharp})(t))$ and $\lim_{t\to\infty} H(f)(t) = H(f_\infty)$ which follows from $f^{\sharp}(t)|\log f^{\sharp}(t)| \leq C\Phi_0(1+|\log\Phi_0|)$ and dominated convergence. □

Remarks. 1. In Theorems 3.1 and 3.2, we do not assume the solutions are small in the norm $\|\cdot\|_\Phi$ of $\mathcal{B}(\Phi)$, and for the entropy equality (3.6) we also do not assume the solutions are positive everywhere.

2. The right-hand-side estimates in (3.4) and (3.7) may also be viewed as the uniqueness of a "final" value problem of (B) (existence results for this problem will be given in the next section). The proofs for this uniqueness are different from those of the uniqueness of the initial value problem; see [L2], [Lu], where the two solutions f, g need not be both strong (say bounded or spatial decay) solutions. This difference is essentially due to the irreversibility of the time evolution of (B), i.e., due to the different collision order $Q^+ - Q^-$ and $Q^- - Q^+$.

4. Two kinds of existence results. In this section the existence of the solutions in Theorems 3.1 and 3.2 will be proven for the initial value problem and for the "final" value problem, respectively. Throughout this section, the collision kernel is assumed to satisfy the same conditions in section 3; i.e., $B(z,\omega)$ satisfies (1.6), (1.7) with $-2 < \gamma \leq 1$, and the functions Φ are given by (1.14) or (1.15), or given by (1.16) for $\gamma \leq 0$, with their constants $k > (3+\gamma)/2,\ a > 0,\ b > 0$, etc. For the initial value problem, we consider the following "interim" equation of nonnegative functions f:

$$f^{\sharp}(x,v,t) = f_0(x,v) + \int_0^t Q(f\wedge\phi, f\wedge\phi)^{\sharp}(x,v,s)ds \quad \text{on} \quad \mathbf{R}^3\times\mathbf{R}^3\times[0,\infty), \tag{4.1}$$

where $f_0 \in \mathcal{B}(\Phi_0)$, $\phi \in \mathcal{B}(\Phi)$ are nonnegative, and

$$(f\wedge\phi)(x,v,t) = \min\{f(x,v,t), \phi(x,v,t)\}.$$

In the following, the positive constants $D^{\pm} = D^{\pm}(B_0,\gamma,k)$ appeared in Lemma 2.5, and (2.19) will be used.

THEOREM 4.1. (i) *If $\gamma < 1$, then for any $0 \leq f_0 \in \mathcal{B}(\Phi_0)$ and any $0 \leq \phi \in \mathcal{B}(\Phi)$,* (4.1) *has a unique nonnegative global solution $f \in \mathcal{B}(\Phi)$. Furthermore, if*

$$\|f_0\|_{\Phi_0} \leq (4D^+\,a\,b^{2+\gamma})^{-1}, \tag{4.2}$$

then for any $\phi \in \mathcal{B}(\Phi)$ *satisfying* $\phi \geq 2\|f_0\|_{\Phi_0}\Phi$, *the corresponding solution* f *is the unique mild solution of* (B) *in* $\mathcal{B}(\Phi)$ *and satisfies* $\|f\|_\Phi \leq 2\|f_0\|_{\Phi_0}$.
(ii) *If* $\gamma = 1$ *and* $0 \leq f_0 \in \mathcal{B}(\Phi_0)$ *satisfying*

$$\|f_0\|_{\Phi_0} \leq (a\,b^3)^{-1}d(1 - dD^+), \qquad 0 < d < [2(D^+ + D^-)]^{-1}, \tag{4.3}$$

then (B) *has a mild solution* $f \in \mathcal{B}(\Phi)$ *satisfying* $\|f\|_\Phi \leq 2\|f_0\|_{\Phi_0}$.

Proof. Consider the operator K_ϕ given by

$$K_\phi(f)(x,v,t) = f_0(x - tv, v) + \int_0^t Q(|f| \wedge \phi, |f| \wedge \phi)^\sharp(x - tv, v, s)ds.$$

By Lemma 2.5 part (iv), K_ϕ maps $C(\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty))$ into $\mathcal{B}(\Phi)$. We first prove that if $f \in \mathcal{B}(\Phi)$ is a fixed point of K_ϕ, then $f \geq 0$ on $\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$ and so f is a global nonnegative solution of (4.1). In addition we have

$$f^\sharp(x,v,t) \leq (\|f_0\|_{\Phi_0} + \|\phi\|_\Phi^2 D^+ a\,b^{2+\gamma})\Phi_0(x,v) \quad \text{on} \quad \mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty). \tag{4.4}$$

The estimate (4.4) follows easily from $f = K_\phi(f)$, Lemma 2.5, and (2.19). Let $\chi(y) = \chi_{[0,\infty)}(y)$ be the characteristic function of $[0,\infty)$. Then, since $f_0 \geq 0$, we have $\forall (x,v) \in \mathbf{R}^3 \times \mathbf{R}^3$

$$\begin{aligned}
(-f^\sharp(x,v,t))^+ &= \int_0^t [-Q(|f| \wedge \phi, |f| \wedge \phi)^\sharp(x,v,s)]\,\chi(-f^\sharp(x,v,s))ds \\
&\leq \int_0^t Q^-(|f| \wedge \phi, |f| \wedge \phi)^\sharp(x,v,s)\,\chi(-f^\sharp(x,v,s))ds \\
&= \int_0^t (|f| \wedge \phi)^\sharp(x,v,s)\,\chi(-f^\sharp(x,v,s))L(|f| \wedge \phi)^\sharp(x,v,s)ds \\
&\leq \|\phi\|_\Phi \int_0^t (-f^\sharp(x,v,s))^+ L(\Phi)^\sharp(x,v,s)ds, \quad t \in [0,\infty).
\end{aligned}$$

Since $L(\Phi)^\sharp(x,v,s)$ is bounded with respect to s (see (2.22)), it follows from the Gronwall inequality that $(-f^\sharp(x,v,t))^+ \equiv 0$, i.e., $f \geq 0$ on $\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$.

(i) $\gamma < 1$. For any $\tau > 0$, define

$$\mathcal{B}(\Phi,\tau) = \{f \in C(\mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)) \mid \|f\|_{\Phi,\tau} := \sup_{t \geq 0} e^{-\tau t}\|f^\sharp(t)\|_{\Phi_0} < \infty\}.$$

Then $\mathcal{B}(\Phi,\tau)$ is a Banach space with the norm $\|f\|_{\Phi,\tau}$. Choose $p = 2$ for $\gamma \leq 0$ and $p = 1/\gamma$ for $0 < \gamma < 1$, and let $q = p/(p-1)$. Then for any $f, g \in \mathcal{B}(\Phi,\tau)$ we have, by $|\,|f| \wedge \phi - |g| \wedge \phi\,| \leq |f - g|$, Lemma 2.5 and (2.18),

$$\begin{aligned}
&|K_\phi(f)^\sharp(x,v,t) - K_\phi(g)^\sharp(x,v,t)| \\
&\leq \int_0^t |\,Q(|f| \wedge \phi, |f| \wedge \phi)^\sharp - Q(|g| \wedge \phi, |g| \wedge \phi)^\sharp\,|(x,v,s)\,ds \\
&\leq 2\|\phi\|_\Phi\,\|f - g\|_{\Phi,\tau} \int_0^t e^{\tau s}[Q^+(\Phi,\Phi)^\sharp + Q^-(\Phi,\Phi)^\sharp](x,v,s)ds \\
&\leq 2\|\phi\|_\Phi\,\|f - g\|_{\Phi,\tau}\,\frac{e^{\tau t}}{(q\tau)^{1/q}}\,(C_p^+ + C_p^-)\Phi_0(x,v), \quad (x,v,t) \in \mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty).
\end{aligned}$$

This implies

$$\|K_\phi(f) - K_\phi(g)\|_{\Phi,\tau} \le \frac{2\|\phi\|_\Phi (C_p^+ + C_p^-)}{(q\tau)^{1/q}} \|f - g\|_{\Phi,\tau}.$$

Let $\tau > 0$ be sufficiently large such that $K_\phi : \mathcal{B}(\Phi,\tau) \to \mathcal{B}(\Phi) \subset \mathcal{B}(\Phi,\tau)$ is contractive. Then K_ϕ has a unique fixed point $f \in \mathcal{B}(\Phi)$, which is therefore the unique nonnegative continuous global solution of (4.1). Now suppose that f_0 satisfies the condition (4.2). Let $0 \le f \in \mathcal{B}(\Phi)$ be the solution of (4.1) corresponding to a "minimum" function $\phi = \underline{\phi} := 2\|f_0\|_{\Phi_0}\Phi$. Then by (4.4) we have $\forall (x,v,t) \in \mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$

$$\begin{aligned} f^\sharp(x,v,t) &\le (\|f_0\|_{\Phi_0} + 4\|f_0\|_{\Phi_0}^2 D^+ a\, b^{2+\gamma})\Phi_0(x,v) \\ &\le 2\|f_0\|_{\Phi_0}\Phi_0(x,v) = \underline{\phi}^\sharp(x,v,t). \end{aligned}$$

Therefore $f \le \underline{\phi}$ and so f is a mild solution of (B). The remainder of the conclusion of part (i) follows easily from the uniqueness of solutions of (4.1).

(ii) $\gamma = 1$. Suppose f_0 satisfies (4.3). Choose $\lambda \in [\,0, d\,]$ such that

$$(a\, b^3)^{-1}\lambda(1 - \lambda D^+) = \|f_0\|_{\Phi_0}. \tag{4.5}$$

Let $\phi = (a\, b^3)^{-1}\lambda\Phi$. For any $f, g \in \mathcal{B}(\Phi)$ we have, as above (using Lemma 2.5 and (2.19)),

$$\|K_\phi(f) - K_\phi(g)\|_\Phi \le 2\lambda(D^+ + D^-)\|f - g\|_\Phi.$$

Because $2\lambda(D^+ + D^-) < 1$, K_ϕ has a unique fixed point $f \in \mathcal{B}(\Phi)$, and $f \ge 0$. Since $\|\phi\|_\Phi = (a\, b^3)^{-1}\lambda$, applying (4.4), (4.5) we obtain $f \le \phi$. Thus f is a mild solution of (B) and, by (4.5), $\|f\|_\Phi \le (a\, b^3)^{-1}\lambda \le 2\|f_0\|_{\Phi_0}$. □

THEOREM 4.2. *For any $F \in C(\mathbf{R}^3 \times \mathbf{R}^3)$ satisfying for $0 < d < [2(D^+ + D^-)]^{-1}$,*

$$(a\, b^{2+\gamma})^{-1} d^2 D^+ \Phi_0 \le F \le (a\, b^{2+\gamma})^{-1} d(1 - dD^-)\Phi_0 \qquad \text{on} \quad \mathbf{R}^3 \times \mathbf{R}^3, \tag{4.6}$$

there exists a mild solution $f \in \mathcal{B}(\Phi)$ of (B) *such that $f_\infty = F$.*

Moreover, if $\gamma < 1$, or $\gamma = 1$ and Φ is given by (1.15) *with $k > 3$ and $\beta = 1$, then the solution f is unique in $\mathcal{B}(\Phi)$.*

Proof. Similar to the existence proof of the initial value problem, in this case we consider the operator $\widetilde{K}_\phi$:

$$\widetilde{K}_\phi(f)(x,v,t) = F(x - tv, v) - \int_t^\infty Q(|f| \wedge \phi, |f| \wedge \phi)^\sharp(x - tv, v, s)ds. \tag{4.7}$$

Let $F \in C(\mathbf{R}^3 \times \mathbf{R}^3)$ satisfy (4.6) and let $\phi = (a\, b^{2+\gamma})^{-1} d\Phi$. Then, by Lemma 2.5, $\widetilde{K}_\phi$ maps $\mathcal{B}(\Phi)$ into $\mathcal{B}(\Phi)$, and as the proof of Theorem 4.1 part (ii) we have

$$\|\widetilde{K}_\phi(f) - \widetilde{K}_\phi(g)\|_\Phi \le 2d(D^+ + D^-)\|f - g\|_\Phi \quad \forall\ f, g \in \mathcal{B}(\Phi).$$

Since $2d(D^+ + D^-) < 1$, $\widetilde{K}_\phi$ has a unique fixed point $f \in \mathcal{B}(\Phi)$. From this and condition (4.6) we have $\forall\ (x,v,t) \in \mathbf{R}^3 \times \mathbf{R}^3 \times [0,\infty)$,

$$\begin{aligned} f^\sharp(x,v,t) &\ge F(x,v) - (a\, b^{2+\gamma})^{-1} d^2 D^+ \Phi_0(x,v) \ge 0, \\ f^\sharp(x,v,t) &\le (a\, b^{2+\gamma})^{-1} d(1 - dD^-)\, \Phi_0(x,v) + (a\, b^{2+\gamma})^{-1} d^2 D^- \Phi_0(x,v) = \phi^\sharp(x,v,t). \end{aligned}$$

Therefore

$$f^\sharp(x,v,t) = F(x,v) - \int_t^\infty Q(f,f)^\sharp(x,v,s)ds \qquad \text{on} \quad \mathbf{R}^3\times\mathbf{R}^3\times[0,\infty) \tag{4.8}$$

and so f is a mild solution of (B) and satisfies $f_\infty = F$ by (3.2) and (3.3). The uniqueness of f in $\mathcal{B}(\Phi)$ follows from Theorems 3.1 and 3.2. □

Remarks. 1. It is easy to show that the class of solutions of Eq.(B) obtained in Theorem 4.1 contains such a subclass of solutions that for $0 \le \gamma \le 1$ the total number of particles is always infinity and for $-2 < \gamma < 0$ the local numbers of particles are always infinity! To see for instance the second case (soft potentials), we use the exponential multiplier form (1.11) to the solutions f and use the estimate (2.19) in Lemma 2.5 for $L(\Phi)^\sharp = Q^-(\Phi,\Phi)^\sharp/\Phi_0$. Then we obtain $f(x,v,t) \ge cf_0(x-tv,v)$ on $\mathbf{R}^3\times\mathbf{R}^3\times[0,\infty)$ where $c = \exp(-\|f\|_\Phi D^- a\, b^{2+\gamma}) > 0$. This implies that if Φ is given by (1.14) or (1.16) with $3/2 \ge k > (3+\gamma)/2$ and if for a small constant $\varepsilon > 0$, the initial data f_0 (in Theorem 4.1) satisfy $f_0 \ge \varepsilon\Phi_0$, then $\int_{\mathbf{R}^3} f(x,v,t)dv \equiv \infty$, $(x,t) \in \mathbf{R}^3\times[0,\infty)$.

For the initial data with infinite total number of particles, as one of their recent results, Mischler and Perthame proved that (see [M,P, Theorem 3.2 for $N = 3$]) if for $q \in (3/2,\ \infty]$ the function $A(z) = \int_{\mathbf{S}^2} B(z,\omega)d\omega$ belongs to $L^q(\mathbf{R}^3)$ and if the initial data f_0 (which need not to be continuous) satisfy $0 \le f_0(x,v) \le \frac{C_0}{6}\exp(-\frac{1}{2}|\frac{x-x_0}{a} - \frac{v-v_0}{b}|^2)$ with $a > 0$, $b > 0$ satisfying $C_0\, a\, b^{\frac{3}{p}-1} < \frac{3-p}{p\|A\|_{L^q(\mathbf{R}^3)}}(\frac{p}{2\pi})^{\frac{3}{2p}}$, $\frac{1}{p} + \frac{1}{q} = 1$, then there exist distributional solutions $f \in L^\infty(\mathbf{R}^3\times\mathbf{R}^3\times(0,\infty))$ of (B) such that $0 \le f(x,v,t) \le \frac{C(t)}{6}\exp(-\frac{1}{2}|\frac{x-tv-x_0}{a} - \frac{v-v_0}{b}|^2)$, where $C(t)$ is bounded on $[0,\infty)$. Comparing this result with our Theorem 3 part (i) (for $-2 < \gamma \le 0$ and for Φ given by (1.16)) one sees that the above condition $A \in L^q(\mathbf{R}^3)$ is too restrictive, and the class of initial data given above is relatively small. However, this was the first global existence result which does not require the initial data decay to zero uniformly in both variables x and v. And our function (1.16), $\Phi_0(x,v) = (1+|\frac{x-x_0}{a} - \frac{v-v_0}{b}|^2)^{-k}$, is referred to this version.

2. In Theorem 4.2, the lower bounds given for F seems too large; it is used only to guarantee the nonnegativity of the solution of the final value problem. A simple example shows that (applying the exponential multiplier form (1.11)) this lower bound cannot be replaced by 0. But we do not know whether this lower bound can be replaced by a small one, for instance, by a traveling Maxwellian $c_1\exp(-c_2|x-x_0|^2 - c_3|v-v_0|^2)$.

3. For soft potentials and the Maxwell model (i.e., $-2 < \gamma \le 0$), a special case of Theorem 4.2 is that the final data $F(x,v)$ is only a function of velocity variable: $F(x,v) = h_1(v - \frac{1}{T}(x-x_0))$, $(T > 0)$. In this case, Theorem 4.2 implies the existence of nonnegative solutions of the future value problem of the spatially homogeneous Boltzmann equation

$$\frac{\partial}{\partial t}h(v,t) = Q(h,h)(v,t), \quad (v,t) \in \mathbf{R}^3\times[0,t_1]; \quad h(v,t_1) = h_1(v), \quad v \in \mathbf{R}^3 \tag{4.9}$$

for certain future data $h_1 > 0$ and for a small $t_1 > 0$, and the corresponding solution f obtained in Theorem 4.2 must be a Nikol'skii's solution (see [N], [T,M, p. 291], or [Ce], [Ko]). That is, f can be written

$$f(x,v,t) = h(v - \frac{1}{T}(x-tv-x_0), Z(t)), \qquad (x,v,t) \in \mathbf{R}^3\times\mathbf{R}^3\times[0,\infty), \tag{4.10}$$

where h is a spatially homogeneous solution on $\mathbf{R}^3 \times [0, t_1]$ with $t_1 = \frac{T}{2+\gamma}$, and

$$Z(t) = \frac{T}{2+\gamma}\left[1 - \left(1 + \frac{t}{T}\right)^{-2-\gamma}\right], \quad t \in [0, \infty) \tag{4.11}$$

To clarify this, let $0 < T < [2(D^+ + D^-)]^{-1}$ and let Φ be given by (1.16) with $a = Tb$, i.e., $\Phi(x, v, t) = \widetilde{\Phi}_0(v - \frac{1}{T}(x - tv - x_0))$, where

$$\widetilde{\Phi}_0(v) = \left(1 + \left|\frac{v - v_0}{b}\right|^2\right)^{-k}, \quad k > (3+\gamma)/2, \quad b > 0.$$

Suppose that $h_1 \in C(\mathbf{R}^3)$ satisfy

$$T(b^{3+\gamma})^{-1} D^+ \widetilde{\Phi}_0(v) \le h_1(v) \le (b^{3+\gamma})^{-1}(1 - TD^-)\widetilde{\Phi}_0(v), \quad v \in \mathbf{R}^3.$$

Let $f \in \mathcal{B}(\Phi)$ be the unique mild solution of (B) obtained in the proof of Theorem 4.2 ($d = T$) corresponding to $f_\infty(x, v) = F(x, v) = h_1(v - \frac{1}{T}(x - x_0))$. By the special form of the final data, we assert first that $f(x, v, t)$ is a function of $(v - \frac{1}{T}(x - tv - x_0), t)$ only. In fact, for any nonnegative function $\widetilde{g} \in C(\mathbf{R}^3 \times [0, \infty))$, let $g(x, v, t) = \widetilde{g}(v - \frac{1}{T}(x - tv - x_0), t)$, then by translation and dilation transforms of velocity variables in collision integrals we have $\forall t \in [0, \infty)$

$$\int_t^\infty Q^\pm(g, g)^\sharp(x, v, s)ds = \int_t^\infty \left(1 + \frac{s}{T}\right)^{-3-\gamma} Q^\pm(\widetilde{g}, \widetilde{g})\left(v - \frac{1}{T}(x - x_0), s\right) ds. \tag{4.12}$$

Thus the operator $\widetilde{K}_\phi$ given by (4) with $F(x - tv, v) = h_1(v - \frac{1}{T}(x - tv - x_0))$ and $\phi(x, v, t) = (b^{3+\gamma})^{-1}\widetilde{\Phi}_0(v - \frac{1}{T}(x - tv - x_0))$ maps the functions of $(v - \frac{1}{T}(x - tv - x_0), t)$ into the same kind of functions. Since $\widetilde{K}_\phi$ is contractive and the solution f is the fixed point of $\widetilde{K}_\phi$, this proves the above assertion. Next, write $f(x, v, t) = \widetilde{f}(v - \frac{1}{T}(x - tv - x_0), t)$. Then $f^\sharp(x, v, t) = \widetilde{f}(v - \frac{1}{T}(x - x_0), t)$. Take $x = x_0$ and replace t by the inverse function of $Z(t)$:

$$Z^{-1}(t) = T\left[\left(1 - \frac{2+\gamma}{T}t\right)^{-1/(2+\gamma)} - 1\right], \qquad t \in [0, t_1), \quad t_1 = T/(2+\gamma).$$

Then by (4.8), (4.12), and change of integral variable s we obtain

$$\widetilde{f}(v, Z^{-1}(t)) = h_1(v) - \int_t^{t_1} Q(\widetilde{f}, \widetilde{f})(v, Z^{-1}(s))ds, \qquad t \in [0, t_1).$$

Moreover, since $f \in \mathcal{B}(\Phi)$, it follows that

$$0 \le \widetilde{f}(v, Z^{-1}(t)) \le C\widetilde{\Phi}_0(v), \quad \lim_{t \to t_1^-} \widetilde{f}(v, Z^{-1}(t)) = h_1(v) \quad \forall v \in \mathbf{R}^3.$$

Therefore, the function $h(v, t) = \widetilde{f}(v, Z^{-1}(t)),\ t \in [0, t_1);\ h(v, t_1) = h_1(v)$, is a classical solution of (4.9), and conversely, f is a Nikol'skii's solution (4.10), (4.11).

Acknowledgments. I would like to thank Prof. Tianquan Chen for valuable discussions on the manuscript. I also wish to thank the referees for their helpful suggestions including pointing out a by-product result at the end of the final Remarks.

REFERENCES

[A,E,P] L. Arkeryd, R. Esposito, and M. Pulvirenti, *The Boltzmann equation for weakly inhomogeneous data,* Comm. Math. Phys., 111 (1987), pp. 393–407.

[B,P,T] N. Bellomo, A. Pelczewski, and G. Toscani, *Mathematical Topics in Nonlinear Kinetic Theory,* World Scientific, Singapore, 1988.

[B,T] N.Bellomo and G.Toscani, *On the Cauchy problem for the nonlinear Boltzmann equation: Global existence, uniqueness and asymptotic behavior,* J. Math. Phys., 26 (1985), pp. 334–338.

[B,B] E. F. Beckenbach and R. Bellman, *Inequalities,* Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1983.

[Ca] T. Carleman, *Problèmes mathématiques dans la théorie cinétique des gaz,* Almqvist & Wiksells, Uppsala, Sweden, 1957.

[Ce] C. Cercignani, *The Boltzmann Equation and Its Applications,* Springer-Verlag, New York, 1988.

[Cs] I. Csiszar, *Information-type measures difference probability distributions and indirect observations,* Studia Sci. Math. Hungar. 2, (1962), pp. 299–318.

[D,L 1] R. J. DiPerna and P. L. Lions, *On the Cauchy problem for Boltzmann equations: Global existence and weak stability,* Ann. of Math. 130, (1989), pp. 321–366.

[D,L 2] R. J. DiPerna and P . L. Lions, *Global solutions of Boltzmann equation and the entropy inequality,* Arch. Rational Mech. Anal., 114 (1991), pp. 47–55.

[Gr] H. Grad, *Asymptotic theory of the Boltzmann equation* II, in Rarefied Gas Dynamics, J. A. Laurmann, ed., Academic Press, New York, 1963, pp. 26–59.

[Gu] T.Gustafsson, *Global $L\hat{p}$-properties for the spatially homogeneous Boltzmann equation,* Arch. Rational Mech. Anal., 103 (1988), pp.1–38.

[H] K. Hamdache, *Existence in the large and asymptotic behavior for the Boltzmann equation,* Japan J. Appl. Math., 2 (1985), pp. 1–15.

[I,S] R. Illner and M. Shinbrot,*The Boltzmann equation: Global existence for a rare gas in an infinite vacuum,* Comm. Math. Phys., 95 (1984), pp. 117–126.

[Ko] M. N. Kogan, *Rarefied Gas Dynamics,* Plenum Press, New York, 1969.

[Ku] S. Kullback, *A lower bound for discrimination information in terms of variation,* IEEE Trans. Inform. Theory, 4 (1967), pp.126–127.

[L1] P. L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications* I, J. Math. Kyoto Univ., 34 (1994), pp. 391–427.

[L2] P. L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications* II, J.Math. Kyoto Univ., 34 (1994), pp. 429–461.

[L3] P. L. Lions, *Compactness in Boltzmann's equation via Fourier integral operators and applications* III,, J. Math. Kyoto Univ., 34 (1994), pp. 539–584.

[Lu] X. G. Lu, *A result on uniqueness of mild solutions of Boltzmann equations,* Transport Theory Statist. Phys., 26 (1997), pp. 209–220.

[M,P] S. Mischler and B. Perthame, *Boltzmann equation with infinite energy: Renormalized solutions and distributional solutions for small initial data close to a Maxwellian,* SIAM J. Math. Anal., 28 (1997), pp. 1015–1027.

[N] A. A. Nikol'skii, *The three-dimensional expansion-contraction of a rarefied gas with power interaction functions,* Soviet Physics-Doklady, 8 (1964), pp. 639–641.

[Pe] B. Perthame, *Time decay, propagation of low moments and dispersive effects for kinetic equations,* Comm. Partial Differential Equations, 21 (1996), pp. 659–686.

[Po] J. Polewczak, *Classical solution of the nonlinear Boltzmann equation in all* R^3*: Asymptotic behavior of solutions,* J. Statist. Phys., 50 (1988), pp. 611–632.

[T1] G. Toscani, *On the Boltzmann equation in unbounded domains,* Arch. Rational Mech. Anal., 95 (1986), pp. 37–49.

[T2] G. Toscani, *H-theorem and asymptotic trend of the solution for a rarefied gas in the vacuum,* Arch. Rational Mech. Anal., 100 (1987), pp. 1–12.

[T,M] C. Truesdell and R. G. Muncaster, *Fundamentals Maxwell's Kinetic Theory of a Simple Monoatomic Gas,* Academic Press, New York, 1980.

[W] B. Wennberg, *Stability and exponential convergence for the Boltzmann equation,* Arch. Rational Mech. Anal., 130 (1995), pp. 103–144.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1175–1184

BOUNDS ON RESONANCES FOR THE LAPLACIAN ON PERTURBATIONS OF HALF-SPACE*

JULIAN EDWARD† AND DAVID PRAVICA‡

Abstract. The resonances of the Laplacian on perturbations of half-spaces of dimensions greater than or equal to two, with either Dirichlet or Neumann boundary conditions, are studied. An upper bound for the resonance counting function is proven. If the domain has an elliptic, nondegenerate, nonglancing periodic billiard trajectory, it is shown that there exists a sequence of resonances that converge to the real axis.

Key words. Laplacian, spectral resonances, resolvent estimates, half-space

AMS subject classifications. 35P25, 35J05, 47F05

PII. S003614109733172X

1. Introduction.

The resonances of the Laplace operator on unbounded manifolds has been the object of considerable study both for exterior domains and for hyperbolic manifolds (see [21] for survey). Less well understood are the resonances for the Laplacian on domains whose boundary extends to infinity.

In this article we study the resonances associated with the Laplacian on compact perturbations of the half-space

$$H^n_+ = \{(x_1, \overline{x}) \in \mathbf{R} \times \mathbf{R}^{n-1},\ x_1 > 0\}.$$

Here we assume $n \geq 2$.

Denote the Laplacian by

$$\Delta = -\frac{\partial^2}{\partial x_1^2} - \cdots - \frac{\partial^2}{\partial x_n^2}.$$

In what follows we will assume either Dirichlet or Neumann boundary conditions.

Let Ω be a connected domain in $\mathbf{R}^n$. Although weaker hypotheses are possible, we assume that the boundary of Ω is a finite, disjoint union of smooth, simple curves and that there exists a positive constant M such that

$$\Omega \cap \{|x| > M\} = H^n_+ \cap \{|x| > M\}.$$

Then by standard methods, for either of the boundary conditions above, the Laplacian can be defined as a self-adjoint operator whose spectrum is given by $\sigma(\Delta) = \sigma_{ac}(\Delta) = [0, \infty)$.

THEOREM 1. *Let $\chi \in C_0^\infty(\mathbf{R}^n)$, and denote its restriction to Ω again by χ. Then the mapping from $\{\Im(k) > 0\}$ to $\mathcal{L}(L^2(\Omega))$ given by*

$$k \to \chi(\Delta - k^2)^{-1}\chi$$

*Received by the editors December 23, 1997; accepted for publication (in revised form) September 10, 1998; published electronically October 4, 1999.

http://www.siam.org/journals/sima/30-6/33172.html

†Department of Mathematics, Florida International University, Miami, FL 33199 (edwardj@fiu.edu).

‡Department of Mathematics, East Carolina University, Greenville, NC (pravica@math.ecu.edu). The research of this author was supported by the ECU Research and Creativity Activity grant.

extends to a meromorphic function in $\mathbf{C}$ *for odd dimensions and to a meromorphic function in* Λ *(the logarithmic plane) in even dimensions.*

Theorem 1 is a simple adaptation of the analogous result in [12], where "black-box" perturbations of $\mathbf{R}^n$ are studied. We remark that the methods of this paper also apply to black-box perturbations of half-space.

We define the resonances of Δ as the poles of $\chi(\Delta - k^2)^{-1}\chi$ and we define the multiplicity of a resonance k_j as the dimension of the image of the projection

$$\frac{i}{\pi}\int_\gamma \chi(\Delta - s^2)^{-1}\chi \; sds$$

with γ a sufficiently small loop around k_j. Let $\{k_j\}_{j\in\mathbf{N}}$ be the list of resonances including multiplicities.

We then prove an upper bound on the number of resonances. Define

$$N(r) = \#\{k_j,\ |k_j| < r\}.$$

For even dimensions, it is more convenient to define

$$N(r,\alpha,\epsilon) = \#\{k_j,\ \epsilon < |k_j| < r,\ |\arg(k_j)| < \alpha\}.$$

THEOREM 2. *For* $n \geq 3$ *odd, there exists a constant* $C > 0$ *such that*

$$N(r) \leq Cr^n.$$

For n *even and for any* $\alpha \in (0,\pi)$, $\epsilon > 0$, *there exists a constant* $C > 0$ *such that*

$$N(r,\alpha,\epsilon) \leq Cr^n.$$

Theorem 2 is proven using the Fredholm determinant method (see [11] and the discussion in section 2). The key estimates on the free resolvent, $(\Delta_{H^n_+} - k^2)^{-1}$, come from observing that the associated Green's function on H^n_+ is determined (via the method of images) by the Green's function on $\mathbf{R}^n$ for either Dirichlet or Neumann boundary conditions.

The restrictions on α, ϵ in even dimensions are not necessary: the arguments in [20] apply in our setting to give a bound on $N(r,\alpha,0)$ which holds $\forall\ \alpha$. However, the proof is much easier for our (weaker) statement, which essentially follows from the proof of the odd dimensional case, after Jensen's formula is replaced by a theorem of Carleman [7]. This fact was kindly pointed out to us by the referee.

If we strengthen our hypotheses about Ω, we can also prove the following.

COROLLARY 1. *Assume that* Ω *satisfies the hypotheses of Theorem* 1*, and furthermore suppose there exists a single, nonglancing, elliptic, nondegenerate trapped ray in* Ω *(see Figure* 1*). Then there exists a sequence of resonances* k'_j *such that* $\Im k'_j \to 0$ *as* $|k'_j| \to \infty$.

Corollary 1 gives an affirmative answer to a conjecture of Lax and Phillips [5] in this setting. We had originally been planning to prove the corollary only in odd dimensions, following the outline of an argument found in [14], [15]. However, the referee kindly suggested an argument that would apply in even dimensions as well and furthermore furnishes lower bounds for the counting function for resonances near the real axis. In proving Corollary 1 we shall adopt the referee's argument.

In our original (odd-dimensional) proof, we began with a global a priori estimate for the cutoff resolvent. Such an estimate was obtained by Zworski [23] for obstacle

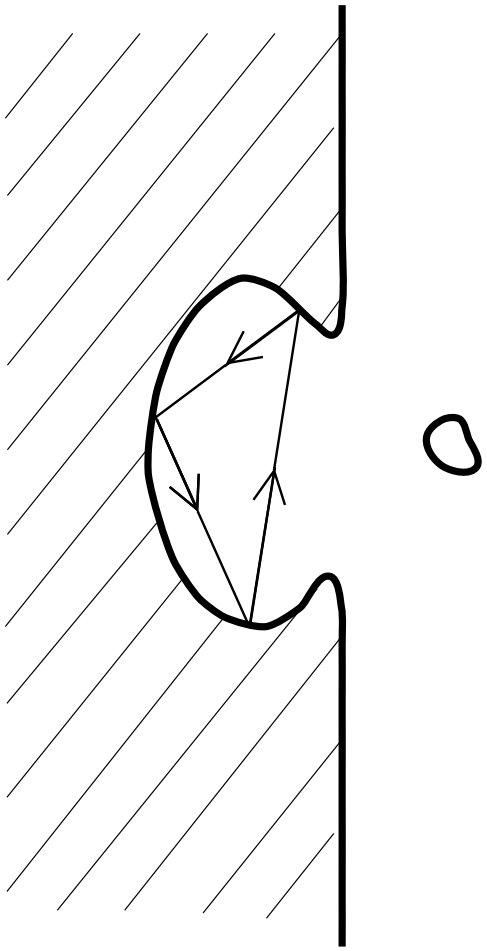


FIG. 1. *Typical domain Ω consisting of a hole in a wall. Dirichlet conditions are imposed on the boundary.*

scattering in odd dimensions, adapting techniques found in [8] and using as a key ingredient the minimodulus theorem for entire functions [17]; the method was then extended to more general perturbations in [14], [15]. Arguing as in [14], [15], we then proved that the existence of quasimodes (see [10], [1], [9], [6]) implies Corollary 1.

The referee pointed out that an a priori bound on the cutoff resolvent can be proven in a neighborhood of the positive real axis near infinity, using the minimodulus theorem of Cartan [7]. Then, using arguments appearing in [16] where they apply to black-box perturbations of $\mathbf{R}^n$, one shows that the existence of quasimodes implies a linear lower bound on the number of resonances converging to the real axis. In [16] it is also noted that under a certain hypothesis on the spacing of the quasimodes, a finer lower bound on the resonances can be obtained.

The arguments of [16] have recently been refined by Stefanov [13], who again studied black-box perturbations of $\mathbf{R}^n$ but without any hypothesis on the spacing of quasimodes. Let $N_q(r)$ be the number of quasimodes less than or equal to r, and let $N_{res}(r)$ be the number of resonances in the set $\{z : \Re z \in [1, r], \Im z \in [0, S(\Re z)]\}$; here

S is a certain function which decays faster than the reciprocal of any polynomial. Then Stefanov proves that for any $k \geq 1$, there exists C_k such that

$$N_{res}(r) \geq N_q(r - r^{-k}) - C_k \ \forall r \geq 1.$$

Stefanov then uses Popov's [9] lower bounds on the counting function for the quasi-modes associated to nondegenerate, elliptic trapped rays whose associated linearized Poincaré map is $2N + 1$ elementary for some half-integer $N \geq 2$, to bound the resonances close to the real axis below by cr^n. The arguments in [13] apply in our setting too, and thus our $O(r^n)$ upper bound on the number of resonances is optimal.

2. Proofs. We will prove the theorems for Dirichlet boundary conditions, and at the end of the section we will state the modifications necessary to treat Neumann conditions.

Denote by $L^2(\Omega)$ the set of square integrable functions on Ω, and denote the set of bounded operators on $L^2(\Omega)$ by $\mathcal{L}(L^2(\Omega))$. Denote by $B(a, r)$ the ball centered at a of radius r. If the related domain is Ω, then $B(a, r)$ refers to the ball intersected with Ω. Denote the Dirichlet Laplacian on Ω (resp., H^n_+) by Δ (resp., Δ_0). Denote the associated resolvent $(\Delta_0 - k^2)^{-1}$ by $R_0(k)$. Denote $(\Delta - k^2)^{-1}$ by $R(k)$. Define the Sobolev spaces $H^i(\Omega)$ as the domains of $(\Delta+1)^{i/2}$. We define a smooth partition of unity $\chi_1 + \chi_2 = 1$ such that $\chi_i \geq 0$, $\text{supp}(\chi_1) \subset B(0, M + 2)$, and $\chi_1 = 1$ on $B(0, M + 1)$. We also define smooth cutoff functions $\tau_i \geq 0$ such that $\tau_1 = 1$ on $\text{supp}(\chi_1)$ and $\text{supp}(\tau_1) \subset B(0, M+3)$, and $\tau_2 = 1$ on $\text{supp}(\chi_2)$ and $\tau_2 = 0$ on $B(0, M)$. Finally, we define a cutoff function $\rho \in C_0^\infty(\mathbf{R}^n)$ such that the $\text{supp}(\rho) \subset B(0, M+4)$ and

$$\rho|_{B(0,M+3)} = 1. \tag{1}$$

LEMMA 1. *Let n be odd (resp., even). Let ψ_1, ψ_2 be smooth functions of bounded support on Ω. Then the mapping from $\{\Im(k) > 0\}$ to $\mathcal{L}(L^2(\Omega))$ given by*

$$k \to \psi_2 R_0(k) \psi_1$$

extends to an entire function on $\mathbf{C}$ (resp., Λ). Also, the same is true for the mapping

$$k \to \frac{\partial}{\partial x_i} \psi_2 R_0(k) \psi_1$$

for $i = 1, \ldots, n$. Furthermore, for $\Im(k) \geq 0$ and $|k| > 1$,

$$\|\psi_2 R_0(k) \psi_1\|_{L^2 \to H^i} \leq C_i (1 + |k|)^{i-1}, \qquad i = 0, 1, 2. \tag{2}$$

Finally, if $\text{supp}(\psi_1) \cap \text{supp}(\psi_2) = \emptyset$, then (2) holds $\forall$ i.

Proof. For $x = (x_1, \overline{x}) \in \mathbf{R} \times \mathbf{R}^{n-1}$ define the mapping $U : \mathbf{R}^n \to \mathbf{R}^n$ by $U(x_1, \overline{x}) = (-x_1, \overline{x})$. Let $\tilde{R}(k) = (\Delta_{\mathbf{R}^n} - k^2)^{-1}$. By the method of images,

$$R_0(k) = \tilde{R}(k) - U^{-1} \tilde{R}(k) U,$$

the right-hand side being restricted to elements of $L^2(H^n_+)$. Hence, dropping the k dependence,

$$\chi_1 R_0 \chi_1 = \chi_1 \tilde{R} \chi_1 - U^{-1} \chi_1 \tilde{R} \chi_1 U.$$

Since U is an isometry on $H^i(\mathbf{R}^n)$, Lemma 1 now follows immediately from the analogous results for $\tilde{R}(k)$ proven in [18] for n odd and [3], [4], [18], [19] for n even. □

Proof of Theorem 1. We prove the result for n odd; the even-dimensional case is a simple adaptation of this. The argument follows closely along the lines of the corresponding result in [12]. We define an approximation of $R(k)$ as follows. Let

$$R_a(k) = \tau_1 R(k_0)\chi_1 + \tau_2 R_0(k)\chi_2. \tag{3}$$

Here k_0 is a parameter to be chosen below.

We have

$$(\Delta - k^2)R_a(k) = I + K \tag{4}$$

with

$$K = (k_0^2 - k^2)\tau_1 R(k_0)\chi_1 + [\Delta, \tau_1]R(k_0)\chi_1 + [\Delta, \tau_2]R_0(k)\chi_2. \tag{5}$$

By (4) we have for $\Im k > 0$

$$R_a = (\Delta - k^2)^{-1}(I + K).$$

By (5) and (1) we have $\rho K = K$; hence

$$R_a\rho = (\Delta - k^2)^{-1}\rho(I + K\rho).$$

For $k = k_0$ and $\operatorname{Im}(k_0) >> 0$, we have by the spectral theorem that $\|K\rho\|_{L^2\to L^2} < 1$ and hence we can write

$$\rho R_a(k)\rho(I + K\rho)^{-1} = \rho R(k)\rho. \tag{6}$$

Fix such a k_0. Now $\rho R_a(k)\rho$ is an entire function of k with values in $\mathcal{L}(L^2(\Omega))$, and therefore meromorphy of $\rho R(k)\rho$ is equivalent to meromorphy of $(I + K\rho)^{-1}$.

On the other hand, since χ_1 and ρ are compactly supported, it follows that $K\rho$ is an entire compact operator-valued function of k. It now follows from the meromorphic Fredholm theorem that $k\to\rho R(k)\rho$ extends to a meromorphic function on $\mathbf{C}$. Finally, it is easy to see that the function ρ can be replaced by any smooth cutoff function. This completes the proof of Theorem 1. □

Proof of Theorem 2. In what follows, let C be various positive constants. A simple Neumann series argument shows that $(I + K\rho)$ is invertible if and only if $(I - (-K\rho)^{n+1})$ is invertible. On the other hand, since $K\rho$ is a pseudodifferential operator of order -1 in $\mathcal{L}(L^2(\Omega))$, with compactly supported Schwartz kernel, it follows that $(K\rho)^{n+1}$ is trace class. Thus the Fredholm determinant $\det(I-(-K\rho)^{n+1})$ is entire on $\mathbf{C}$ for n odd and on Λ for n even. Furthermore, the following lemma applies.

LEMMA 2. *The resonances of* Δ *(counted with their multiplicities) are among the zeros of the function*

$$k\to h(k) \equiv \det(I - (-K\rho)^{n+1}(k)),$$

counted with their multiplicities.

The reader is referred to [20] for a proof of this result.

We shall now complete the proof of Theorem 2 for n odd, after which the modifications necessary for n even will be indicated.

Because of Lemma 2, the bound appearing in Theorem 2 will follow from Jensen's inequality and the estimate

$$|h(k)| \leq C \exp C|k|^n. \tag{7}$$

To obtain this estimate, we first write $K\rho = K_1 + K_2$ with $K_2 = [\Delta, \tau_2]R_0(k)\chi_2\rho$. We then apply the theory of characteristic values developed in [2] and adapted to exterior problems in [8], [22], [18]. The characteristic values $\mu_j(A)$ of a compact operator A are the eigenvalues, listed in increasing order and counting multiplicities, of the operator $|A|$. We recall the following inequalities from [2]: $\mu_{j+k-1}(AB) \leq \mu_j(A)\mu_k(B)$, $\mu_{j+k-1}(A+B) \leq \mu_j(A) + \mu_k(B), and\ \mu_j(AB) \leq \|A\|\mu_j(B)$.

Applying inequalities on Fredholm determinants appearing in [2], we get

$$\begin{aligned}\det(I - (-K\rho)^{n+1}) &\leq \det(I + 2^n|K_1|^{n+1})^{2n+2}\ \det(I + 2^n|K_2|^{n+1})^{2n+2} \\ &\leq \left(\prod_{j=1}^{\infty}(1 + 2^n\mu_j(|K_1|)^{n+1})\right)^{2n+2}\left(\prod_{j=1}^{\infty}(1 + 2^n\mu_j(|K_2|)^{n+1})\right)^{2n+2}. \end{aligned} \tag{8}$$

We estimate first the term involving K_1. Recall

$$K_1 = (k^2 - k_0^2)\tau_1 R(k_0)\chi_1\rho + [\Delta, \tau_1]R(k_0)\chi_1\rho.$$

Since τ_1, χ_1 are compactly supported, it follows by standard eigenvalue asymptotics for pseudodifferential operators [11] that

$$\mu_j(|\tau_1 R(k_0)\chi_1\rho|) \sim Cj^{-2/n}$$

and

$$\mu_j(|[\Delta, \tau_1]R(k_0)\chi_1\rho|) \sim Cj^{-1/n}.$$

It follows that, denoting the largest integer below x by $\lfloor x \rfloor$,

$$\mu_{j-1}(|K_1|) \leq C|k|^2\lfloor j/2 \rfloor^{-2/n} + C\lfloor j/2 \rfloor^{-1/n}.$$

Hence we get

$$\begin{aligned}\mu_j(|K_1|^{n+1}) &\leq (\mu_{(\lfloor j/(n+1)\rfloor+1)}(|K_1|))^{n+1} \\ &\leq (C|k|^2(\lfloor j/2(n+1) + 2\rfloor^{-2/n} + C(\lfloor j/2(n+1)\rfloor + 2)^{-1/n})^{n+1} \\ &\leq (C|k|^2\lfloor j/2(n+1)\rfloor^{-2/n} + C\lfloor j/2(n+1)\rfloor^{-1/n})^{n+1} \\ &\leq C|k|^{2n+2}j^{-(2n+2)/n} + Cj^{-1-1/n}.\end{aligned}$$

We have, for $|k|$ sufficiently large,

$$\begin{aligned}&\prod_{j=1}^{\infty}(1 + 2^n\mu_j(|K_1|)^{n+1}) \\ &\quad\leq \prod_{j\leq k^{2n}}(1 + C|k/j^{1/n}|^{2n+2}) \prod_{j>k^{2n}}(1 + C|j^{-1/2n}|^{2n+2} + Cj^{-1-1/n}).\end{aligned}$$

These two factors are bounded as in [22]; we sketch the argument. The first factor is bounded by comparing it to

$$\exp\left(\int_1^{|k|^2} \ln(1 + C|k/x^{1/n}|^{2n+2})dx\right),$$

which is bounded by $e^{C|k|^n}$. The second factor is treated similarly. Thus

$$\prod_{j=1}^{\infty}(1+2^n\mu_j(|K_1|)^{n+1}) \le e^{C|k|^n}. \tag{9}$$

Note that this estimate holds for all $k \in \mathbf{C}$.

Next, we estimate the terms involving K_2. The argument sketched below is an adaptation of one originally implemented by Vodev in [18].

LEMMA 3. *Suppose* $\Im k \ge 0$. *Then* $\det(I+2^n|K_2|^{n+1}) < C$ *for some positive constant* C.

Proof. We have, using (2) with $i=3$,

$$\begin{aligned}\mu_j(K_2) &= \mu_j(\rho(I+\Delta)^{-1}(I+\Delta)K_2)\\ &\le \mu_j(\rho(I+\Delta)^{-1})\|(I+\Delta)K_2)\|\\ &\le Cj^{-2/n}|k|^2.\end{aligned}$$

Now the arguments leading to (9) are easily adapted to this case. The details are omitted.

LEMMA 4. *Suppose* $\Im k < 0$. *Then*

$$\det(I+2^n|K_2|^{n+1}) \le \exp(C|k|^n).$$

Proof. We use the notation of Lemma 1. We first observe that for $\Im k < 0$,

$$\begin{aligned}&\det(I+2^n|K_2(k)|^{n+1})\\ &\quad\le \det(I+2^n|K_2(-k)|^{n+1})^{2n+2}\det(I+2^n|K_2(k)-K_2(-k)|^{n+1})^{2n+2}.\end{aligned} \tag{10}$$

The first term on the right-hand side is bounded by Lemma 3. To bound the second, we begin by recalling Stone's theorem for the Laplacian on $\mathbf{R}^n$,

$$\tilde{R}(k)-\tilde{R}(-k) = c_n k^{n-2}\int_{S^{n-1}} e^{ik<\omega,x-y>}d\omega.$$

It follows that

$$(R_0(k)-R_0(-k))(x,y) = c_n k^{n-2}\int_{S^{n-1}} e^{ik\langle\omega,x-y\rangle} - e^{-ik\omega_1(x_1-y_1)}e^{ik\langle\overline{\omega},\overline{x}-\overline{y}\rangle}d\omega.$$

This allows us the factorization for $\Im k > 0$

$$\rho(R_0(k)-R_0(-k))\rho = c_n k^{n-2}\rho E^*(k)E(k)\rho,$$

with $E(k)$ being a mapping from $L^2(H^n_+)$ to $L^2(S^{n-1})$ whose Schwartz kernel is given by

$$e^{-ik\langle\omega,y\rangle} - e^{ik\omega_1 y_1}e^{-ik\langle\overline{\omega},\overline{y}\rangle}.$$

Observe that this kernel is an entire function in k, and we have the estimate

$$\|\rho E^*(k)\|_{L^2\to L^2} \le e^{(M+2)|\Im k|},$$

with M as in (1). Also, observe that

$$\begin{aligned}\mu_j(E(k)\rho) &\le \mu_j((\Delta_{S^{n-1}}+1)^{-m})\|(\Delta_{S^{n-1}}+1)^m E(k)\rho\|\\ &\le C^m(j)^{-2m/(n-1)}\|(\Delta_{S^{n-1}}+1)^m E(k)\rho\|.\end{aligned}$$

On the other hand, the estimate $|k|^p e^{C|k|} \leq Cp!\ e^{(C+1)|k|}$ together with a combinatorial argument yield

$$\|(\Delta_{S^{n-1}} + 1)^m E(k)\rho\| \leq (2m)!\ e^{C|k|}.$$

Thus

$$\mu_j(E(k)\rho) \leq (2m)!\ C^m (j)^{-2m/(n-1)} e^{C|k|}.$$

Optimizing over m, we get

$$\mu_j(E(k)\rho) \leq e^{-(1/j^{(n-1)})/C} e^{C|k|}.$$

Hence

$$\mu_j(\rho(R_0(k) - R_0(-k))\rho) \leq e^{-(1/j^{(n-1)})/C} e^{C|k|}.$$

A similar argument yields

$$\mu_j\left(\rho \frac{\partial}{\partial x_i}(R_0(k) - R_0(-k))\rho\right) \leq e^{-(j^{1/(n-1)})/C} e^{C|k|}, \qquad i = 1, \ldots, n.$$

Combining these yields for $\Im k < 0$,

$$\mu_j\left([\Delta, \tau_2](R_0(k) - R_0(-k))\chi_2\rho\right) \leq e^{C|k|} e^{-(j^{1/(n-1)})/C}.$$

Thus, setting $T(k) = K_2(k) - K_2(-k)$,

$$\begin{aligned}
\det(I + 2^n |T(k)|^{n+1}) &\leq \prod_{j < C|k|^{n-1}} (1 + 2^n \mu_j(|T(k)|^{n+1})) \\
&\qquad \times \exp\left(\sum_{j \geq C|k|^{n-1}} 2^n \mu_j(|T(k)|^{n+1})\right) \\
&\leq \prod_{j < C|k|^{n-1}} e^{C|k|} \exp\left(\sum_{j \geq C|k|^{n-1}} \left(e^{C|k|} e^{-(j^{1/(n-1)})/C}\right)^{n+1}\right) \\
&\leq e^{C|k|^n}.
\end{aligned}$$

This completes the proof of the lemma. □

Combining (8) and (9) and Lemmas 3 and 4, we obtain (7) and hence Theorem 2 for n odd.

We now complete the proof for n even. In place of Jensen's formula, we use Carleman's theorem [7, Thm. 5.1.1]. The bounds on the determinant for large k are proven as in the odd-dimensional case (set the branch line on the negative real axis in Lemma 4), and the bounds for k near the origin are obtained simply by analyticity.

Proof of Corollary 1.

LEMMA 5. *Let* $\Lambda_{\alpha,N} = \{k : |\arg k| < \alpha,\ |k| > N\}$. *Fix* $\epsilon > 0$, $\alpha \in (0, \pi/2)$. *Let* $\{k_j\}$ *denote the resonances in* $\Lambda_{\alpha,N}$. *Then there exist constants* N, C *such that*

$$\|\rho R(k)\rho\| \leq C e^{C|k|^n \log k} \tag{11}$$

$\forall\ k \in \Lambda_{\alpha,N} - \cup_{k_j} B(k_j, |k_j|^{-n-\epsilon})$.

Proof. In what follows we assume $|k| > 1$. We recall

$$\rho R_a(k)\rho(I + K\rho)^{-1} = \rho R(k)\rho. \tag{12}$$

It follows from (3), Lemma 1, and the proof of Lemma 4 that for $\alpha \in (0, \pi)$ and $k \in \Lambda_\alpha$,

$$\|\rho R_a \rho\|_{L^2 \to L^2} \leq e^{C|k|^n}. \tag{13}$$

To bound $(1 + K\rho)^{-1}$, we proceed as follows: From [2, Chap. 5, Thm. 5.1.], we have

$$\|(I + K\rho)^{-1}\|_{L^2 \to L^2} \leq |\det(I + (K\rho)^{n+1})|^{-1} \det(I + |K\rho|^{n+1})^{n+1}.$$

By the proof of Theorem 2, we have

$$\det(I + |K\rho|^{n+1})^{n+1} \leq e^{C|k|^n}. \tag{14}$$

To obtain the lower bound on $f(k) \equiv |\det(I + (K\rho)^{n+1})|$, we first note that

$$\begin{aligned} |\det(I + (K\rho)^{n+1})|^{n+1} &\leq \det(I + |K\rho|^{n+1})^{n+1} \\ &\leq e^{C|k|^n}. \end{aligned}$$

Next, we apply the minimodulus theorem of Cartan (see [7, p. 21]). Thus let $\eta, c > 0$. Suppose $l >> 0$. Clearly, $|f(k)| \leq e^{Cl^n}$ for $k \in B(l, 2ecl)$. It follows that in $B(l, cl)$, and outside a system of disks of radii whose sum is no greater than $4cl\eta$, we have

$$|f(k)| > (e^{C|k|^n})^{-2-\log(3e/2\eta)}, \tag{15}$$

with C a constant independent of l, k. Setting $\eta = l^{-n-2}$ and combining the inequality above with (14), (15), (13), and (12), the inequality appearing in (11) holds in $B(l, cl)$ in the complement of the system of disks. We decompose the system of disks into the union $\cup U_j$, where U_j are connected, mutually disjoint, and have diameter at most $4cl^{-n-1}$. By increasing c slightly if necessary, we also have that $U_j \subset B(l, cl)$. Finally, we can assume that each U_j contains a resonance. For if not, then (11) holds on U_j by the maximum principle.

Now we suppose l is sufficiently large that $l^{-n-\epsilon} > 8cl^{-n-1}$. Then it follows that for each j, $U_j \subset B(k_i, |k_i|^{-n-\epsilon})$ for some resonance $k_i \in B(l, cl)$. With α, N determined by c, ϵ, the lemma now follows. □

The proof of Corollary 1 is completed as follows: applying Lemma 5 of this paper, along with Lemma 2 of [16] (with $h = 1/k^2$), the proofs of Theorems 1 and 2 in [16] carry over immediately, yielding a linear lower bound on the resonances converging to the real axis.

In the case of Neumann boundary conditions, it suffices to observe that by the method of images, the associated resolvent of half-space satisfies

$$R_0(k) = \tilde{R}(k) + U^{-1}\tilde{R}(k)U.$$

The preceding arguments carry over without difficulty.

Acknowledgments. The first author thanks Maciej Zworski for guiding him through the intricacies of scattering theory. The second author would like to thank ECU for providing a summer research grant. We also thank the referee, whose suggestions improved both the presentation and the content of this paper.

REFERENCES

[1] V.M. Babich and V.S. Buldyrev, *Short-Wavelength Diffraction Theory. Asymptotic Methods*, Springer Ser. Wave Phenomena 4, Springer-Verlag, Berlin, 1991.

[2] I. Gohberg and M. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr., AMS, Providence, RI, 1969.

[3] A. Intissar, *A polynomial bound on the number of scattering poles for a potential in even dimensional space in* $\mathbf{R}^n$, Comm. Partial Differential Equations, 11 (1986), pp. 367–396.

[4] A. Intissar, *On the Value Distribution of the Scattering Poles Associated to the Schrödinger Operator* $H + (-i\nabla + b(x))^2 + a(x)$ *in* $\mathbf{R}^n$, $n \geq 3$, unpublished.

[5] P. Lax and R. Phillips, *Scattering Theory*, Academic Press, New York, 1989.

[6] V.F. Lazutkin, *KAM Theory and Semiclassical Approximations to Eigenfunctions*, addendum by A.I. Schnirelman., Ergeb. Math. Grenzgeb. (3) 24, Springer-Verlag, Berlin, 1993.

[7] B.Ja. Levin, *Distribution of Zeros of Entire Functions*, Transl. Math. Monogr. 5, AMS, Providence, RI, 1964.

[8] R.B. Melrose, *Polynomial bounds on the distribution of poles in scattering by an obstacle*, in Journées Equations aux derivées partielles, Saint-Jean-de-Monts, France, 1984.

[9] G. Popov, *Quasimodes for the Laplace operator and glancing hypersurfaces*, in Microlocal Analysis and Nonlinear Waves, M. Beals, R. Melrose, J. Rauch, eds., Springer-Verlag, Berlin, New York, 1991.

[10] J. Ralston, *Approximate eigenfunctions of the Laplacian*, J. Differential Geom., 12 (1977), pp. 87–100.

[11] M. Shubin, *Pseudodifferential Operators and Spectral Theory*, Springer-Verlag, Berlin, New York, 1987.

[12] J. Sjostrand and M. Zworski, *Complex scaling and the distribution of scattering poles*, J. Amer. Math. Soc., 4 (1991), pp. 729–769.

[13] P. Stefanov, *Quasimodes and resonances: Fine lower bounds*, Duke Math. J., 99 (1999), pp. 75–92.

[14] P. Stefanov and G. Vodev, *Distribution of resonances for Neumann problem in linear elasticity outside a strictly convex body*, Duke Math. J., 78 (1995), pp. 677–714.

[15] P. Stefanov and G. Vodev, *Neumann resonances in linear elasticity for an arbitrary body*, Comm. Math. Phys., 176 (1996), pp. 645–659.

[16] S. Tang and M. Zworski, *From quasimodes to resonances*, Math. Res. Lett., 5 (1998), pp. 261–272.

[17] E. Titchmarsh, *The Theory of Functions*, Oxford University Press, Oxford, UK, 1939.

[18] G. Vodev, *Sharp polynomial bounds on the number of scattering poles for perturbations of the Laplacian*, Comm. Math. Phys., 146 (1992), pp. 205–216.

[19] G. Vodev, *Sharp bounds on the number of scattering poles in the two dimensional case*, Math. Nachr., 170 (1994), pp. 287–297.

[20] G. Vodev, *Sharp bounds on the number of scattering poles in even-dimensional spaces*, Duke Math. J., 74 (1994), pp. 1–16.

[21] M. Zworski, *Counting the scattering poles*, in Spectral and Scattering Theory, M. Ikawa, ed., Marcel Dekker, New York, Basel, 1993.

[22] M. Zworski, *Sharp polynomial bounds on the number of scattering poles*, Duke. Math. J., 59 (1989), pp. 311–323.

[23] M. Zworski, unpublished, 1990.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1185–1207

ON A CONJECTURE RELATIVE TO THE MAXIMA OF HARMONIC FUNCTIONS ON CONVEX DOMAINS*

LUCIO R. BERRONE†

Abstract. We consider a harmonic function u defined on a bounded domain $\Omega \subset \mathbb{R}^2$ and satisfying the mixed boundary conditions $u|_{\Gamma_0} = 0$, $(\partial u/\partial n)|_{\Gamma_1} = 1$, where Γ_1 is composed by a finite number of arcs of $\partial\Omega$ and $\Gamma_0 = \partial\Omega \sim \overline{\Gamma_1}$. In [Berrone, *Subsistencia de Modelos Matematicos que Involucran a la Ecuacion del Calor-Difusion*, Ph.D. thesis, Universidad Nacional de Rosario, Argentina, 1994] it was conjectured that if Ω is convex and the subset Γ_1 is made to vary on $\partial\Omega$ so as to maintain its measure equal to a constant $C > 0$, then $\Gamma_1 \mapsto \sup_{x\in\Omega} u$ attains its maximum value when Γ_1 is a certain *connected* arc of measure C. The present paper has evolved from attempts to prove this conjecture. When certain geometric restrictions are satisfied by the components of Γ_1, the property stated by the conjecture is shown to hold for every regular domain Ω, convex or not, and *every* connected arc, provided that the measure $|\Gamma_1|$ is sufficiently small (see Theorem 5). However, convexity becomes a necessary condition in order that the full conjecture can be supportable (see section 2). In addition, some variations of the conjecture are proposed.

Key words. harmonic functions, mixed boundary value problems

AMS subject classifications. 35J05, 35B99

PII. S0036141098334973

1. Introduction.

Let Ω be a bounded plane domain whose boundary curve $\partial\Omega$ is composed of two families of relatively open arcs Γ_0 and Γ_1 such that $\partial\Omega = \overline{\Gamma_0 \cup \Gamma_1}$, $\Gamma_0 \cap \Gamma_1 = \emptyset$. In this paper we are concerned with the following mixed boundary value problem:

$$\left|\begin{array}{ll} \Delta u(x) = 0, & x \in \Omega, \\ u(x) = 0, & x \in \Gamma_0, \\ \frac{\partial u}{\partial n}(x) = 1, & x \in \Gamma_1. \end{array}\right. \tag{1}$$

In (1), we have denoted by n the unit outward-pointing normal vector to $\partial\Omega$. In regard to the required regularity of Ω, a Dini-smooth boundary $\partial\Omega$ will be assumed in the developments of sections 3 and 4. Moreover, an interior sphere condition on every point of $\partial\Omega$ will be supposed in order that the strong maximum principle and Hopf's lemma may hold. As is well known, the solution to problem (1) is continuous up to the boundary provided that the family Γ_1 is finite, as we shall assume henceforth. This solution will be denoted by $u[\Gamma_1]$ to indicate its dependence on Γ_1. It is also well known that many physical phenomena are modeled by problem (1). A nonexhaustive list of these is given in [15] (see also [14]) but, for our purpose, it will be illustrative to think about problem (1) as giving the equilibrium position of an elastic membrane Ω which is subjected to a unit normal force on Γ_1, while it is fixed at level zero along the remaining portion Γ_0 of the boundary of Ω.

Our main interest focuses on the behavior of the functional $\Gamma_1 \mapsto \sup_{x\in\Omega} u[\Gamma_1]$. Concretely, in the thesis [2] the following conjecture has been posed.

*Received by the editors March 6, 1998; accepted for publication (in revised form) January 11, 1999; published electronically October 4, 1999. This work was supported by a Beca Externa from "Consejo Nacional de Investigaciones Científicas y Técnicas" de la República Argentina.
http://www.siam.org/journals/sima/30-6/33497.html
†Departamento de Matemática, Av. Pellegrini 250, 2000 Rosario, Argentina (berrone@unrctu.edu.ar).

CONJECTURE 1. *Let Ω be a convex domain and C be a constant such that $0 < C < |\partial\Omega|$. If Γ_1 is made to vary on all finite families of arcs of $\partial\Omega$ with measure $|\Gamma_1| = C$, then $\Gamma_1 \mapsto \sup_{x\in\Omega} u[\Gamma_1]$ reaches its maximum value when Γ_1 is a certain* connected *arc of $\partial\Omega$.*

In terms of the above-given mechanical interpretation of problem (1), the conjecture states that if a convex membrane initially fixed at zero is lifted by a unitary normal force on portions Γ_1 of its boundary with a constant total measure C, then the membrane will reach a maximum height when Γ_1 is a certain connected arc of measure C. As we explained in [4], this conjecture arose from attempts to estimate the solution to boundary value problems like (1) through sub- and supersolutions: in general, simple sub- and supersolutions to problem (1) are more easily calculated when Γ_1 is connected (cf. [3]). Furthermore, exact solutions to (1) are known for particular domains in this situation, Ghizzetti's solution for the circle (see [7]) being a good example. Ghizzeti's solution will be of capital importance in the developments of sections 3 and 4.

In [4], Conjecture 1 is studied for unbounded domains and it is shown there to be not generally true in this case. For instance, Conjecture 1 holds for the half-plane but it fails for an infinite strip. Section 2 of this paper is devoted to supporting the hypothesis of convexity of the domain Ω in the conjecture by showing an example of a nonconvex domain such that connected arcs of its boundary are not generally optimal for $\Gamma_1 \mapsto \sup_{x\in\Omega} u[\Gamma_1]$.

Up to now, we have been unable generally to prove or to construct a counterexample to Conjecture 1. Nevertheless, some variations of the conjecture seem to be more manageable. An interesting variation is obtained when the functional $\Gamma_1 \mapsto \|u[\Gamma_1]\|_p = \left(\int_{\partial\Omega} |u[\Gamma_1]|^p \, ds\right)^{1/p}$, $1 \le p < +\infty$, is taken instead of $\Gamma_1 \mapsto \sup_{x\in\Omega} u[\Gamma_1]$ in Conjecture 1, giving rise to the following.

CONJECTURE 2. *Let Ω be a convex domain and C be a constant such that $0 < C < |\partial\Omega|$. If Γ_1 is made to vary on all possible finite families of arcs of $\partial\Omega$ with total measure $|\Gamma_1| = C$, then $\Gamma_1 \mapsto \left(\int_{\partial\Omega} |u[\Gamma_1]|^p \, ds\right)^{1/p}$ reaches its maximum value when Γ_1 is a certain* connected *arc of $\partial\Omega$.*

In mechanical terms, Conjecture 2 for $p = 1$ asserts that in order to maximize the mean height of the boundary of the membrane by lifting a portion of measure C of its boundary, a connected portion of measure C must be lifted. Now, if the Green formula $\int_\Omega u\,\Delta u\,dx = \int_{\partial\Omega} u\frac{\partial u}{\partial n} ds - \int_\Omega |\nabla u|^2\,dx$ is applied to the solution $u[\Gamma_1]$ to problem (1), we obtain

$$\int_{\partial\Omega} u[\Gamma_1]\,ds = \int_{\Gamma_1} u[\Gamma_1]\,ds = \int_\Omega |\nabla u[\Gamma_1]|^2\,dx,$$

so that maximizing $\int_{\partial\Omega} u[\Gamma_1]\,ds$ amounts to the same as maximizing the Dirichlet integral $\int_\Omega |\nabla u[\Gamma_1]|^2\,dx$. On the other hand, there exists a point $P \in \Omega$ such that $|\partial\Omega|\,u[\Gamma_1](P) = \int_{\partial\Omega} u[\Gamma_1]\,ds$ and then, the following variation of Conjecture 2 is suggested which says that the height of the membrane at the interior point P is a maximum when Γ_1 is connected.

CONJECTURE 3. *Let Ω be a convex domain and fix $P \in \Omega$. Moreover, let C be a constant such that $0 < C < |\partial\Omega|$. If Γ_1 is made to vary on all possible finite families of arcs of $\partial\Omega$ with total measure $|\Gamma_1| = C$, then $u[\Gamma_1](P)$ reaches its maximum value when Γ_1 is a certain* connected *subarc of $\partial\Omega$.*

Observe that if Ω is a circle of radius R centered at the origin, we have $\int_{\partial\Omega} u[\Gamma_1]\,ds = 2\pi R\,u[\Gamma_1](0)$ and then Conjecture 2 for $p = 1$ and Conjecture 3 with $P = 0$ are equiv-

alent statements. Recently, a proof of this particular instance was obtained (see [5]) by relating it to a nice result of Beurling on capacities of subsets of the circumference.

Of course, many other variations and extensions of these conjectures are viable but in this paper we discuss two concrete results related to Conjecture 1. To properly state these results, first we need to introduce some preliminary concepts and notations. Let us consider in the following a Jordan domain Ω with a sufficiently regular boundary curve γ. We suppose that γ is parametrized by its arc length s measured from some point $O \in \gamma$. A relatively open subset Γ_1 of γ with a finite number of components can be described as follows:

$$\Gamma_1 = \{\gamma(s) : a_k < s < b_k,\ k = 1, 2, \ldots, n\}, \tag{2}$$

but, for the sake of brevity, we will identify a point $\gamma(s)$ belonging to γ with its corresponding arc length s, so that the more condensed notations

$$\Gamma_1 = \cup_{k=1}^{n}(a_k, b_k) = \cup_{k=1}^{n}\Gamma_1^{(k)}, \tag{3}$$

$$\Gamma_1^{(k)} = (a_k, b_k),\ k = 1, 2, \ldots, n \tag{4}$$

will be used throughout this paper to denote the family of arcs given by (2). We indicate by Γ_1^* a generic connected arc of γ, i.e.,

$$\Gamma_1^* = (a, b). \tag{5}$$

Furthermore, if Γ_1 is given by (3), for $0 < \varepsilon \leq 1$ we define

$$\Gamma_1(\varepsilon) = \cup_{k=1}^{n}\left(a_k(\varepsilon), b_k(\varepsilon)\right), \tag{6}$$

where

$$a_k(\varepsilon) = \frac{a_k + b_k}{2} - \frac{\varepsilon}{2}\left(b_k - a_k\right)$$

and

$$b_k(\varepsilon) = \frac{a_k + b_k}{2} + \frac{\varepsilon}{2}\left(b_k - a_k\right).$$

Note that $\Gamma_1(\varepsilon)$ corresponds to a shrinkage of Γ_1 of magnitude ε and therefore, $|\Gamma_1(\varepsilon)| = \varepsilon\ |\Gamma_1|$.

A proof is given in section 3 of the following result.

THEOREM 4. *Let Ω be a Jordan domain with a Dini-smooth boundary γ and satisfying an interior sphere condition at every point of γ. Moreover, let Γ_1 and Γ_1^* be two subsets of γ respectively given by* (3) *and* (5) *with $b - a = \sum_{k=1}^{n}(b_k - a_k)$. Then, the inequality*

$$\sup_{\Omega} u[\Gamma_1(\varepsilon)] < \sup_{\Omega} u[\Gamma_1^*(\varepsilon)] \tag{7}$$

holds for every sufficiently small ε.

Recall that a uniformly continuous function φ defined on a connected set $A \subset \mathbb{C}$ is said to be Dini-continuous when

$$\int_0^{\delta} \frac{\omega(\varphi, t)}{t}\, dt < +\infty,$$

where $\omega(\varphi,t)=\sup\{|\varphi(z_1)-\varphi(z_2)|:z_1,\,z_2\in A,\ |z_1-z_2|\le t\}$ is the modulus of continuity of φ and $\delta>0$. If $\partial\Omega$ admits a parametrization $\gamma(t)$, $a\le t\le b$, such that γ' is Dini-continuous and $\gamma'(t)\neq 0$, $a\le t\le b$, then we say the domain Ω has a Dini-smooth boundary. This regularity condition on $\partial\Omega$ suffices for a conformal map $f:B_1(0)\to\Omega$ to have a derivative f' which can be continuously extended to $\overline{\Omega}$ (cf. [12, Theorem 3.5, p. 48]), a fact we exploit often in sections 3 and 4.

For a finite family of arcs $\Gamma_1\subset\partial\Omega$, we define the quantity $d=d(\Gamma_1)$ to be the minimum distance in $\partial\Omega$ existing between adjacent components of Γ_1. Thus, if Γ_1 is represented by (3), then we set

$$d(\Gamma_1)=\min\left\{\min_{1\le k\le n-1}\left(a_{k+1}-b_k\right),a_1-b_n\right\}. \tag{8}$$

The quantities

$$H(\Gamma_1)=\frac{|\Gamma_1|}{\sin\left[d(\Gamma_1)/\left(2\left\|f'\right\|_\infty\right)\right]} \tag{9}$$

and

$$\rho(\Gamma_1)=\frac{1}{|\Gamma_1|}\max_{1\le k\le n}\left|\Gamma_1^{(k)}\right|, \tag{10}$$

where f is a conformal mapping of $B_1(0)$ onto Ω, are shown to be useful in the developments of section 4. We remark that $\left\|f'\right\|_\infty=\sup_{B_1(0)}|f'|$ does not depend on the particular choice of the mapping f but only on the geometry of Ω. The same property, that is, independent of the chosen mapping f, is enjoyed by the quantity $M(f)=\inf_{\partial B_1(0)}|f'|$. Taking into account that $|\partial\Omega|=\int_0^{2\pi}\left|f'(e^{i\theta})\right|d\theta\le 2\pi\left\|f'\right\|_\infty$ and $d/|\partial\Omega|<1/2$, we see that $0<d/\left(2\left\|f'\right\|_\infty\right)\le\pi d/|\partial\Omega|<\pi/2$. Therefore, $H(\Gamma_1)$ is a decreasing function of $d(\Gamma_1)$. It should also be noted that $nd\le|\Gamma_0|$; hence the number of components n is bounded by $|\Gamma_0|/d$.

Using the quantities $H(\Gamma_1)$ and $\rho(\Gamma_1)$, a distinguished class $\mathcal{F}$ of finite families of arcs of $\partial\Omega$ is now defined as follows:

$$\mathcal{F}=\left\{\Gamma_1\subset\partial\Omega:\rho(\Gamma_1)\le\delta_1,\ \frac{\left\|f'\right\|_\infty}{8\pi\,M(f')}H^2(\Gamma_1)\le 1-\delta_2,\ 0<\delta_1<\delta_2<1\right\}.$$

Section 4 is devoted to proving the following generalization of Theorem 4.

THEOREM 5. *Let Ω be a Jordan domain with a Dini-smooth boundary γ and satisfying an interior sphere condition at every point of γ. Then, for every family $\Gamma_1\in\mathcal{F}$* with sufficiently small measure*, the inequality*

$$\sup_\Omega u[\Gamma_1]<\sup_\Omega u[\Gamma_1^*]$$

holds provided that $\Gamma_1^*\subset\partial\Omega$ is a connected arc *with $|\Gamma_1^*|=|\Gamma_1|$.*

For the shrinkage $\Gamma_1(\varepsilon)$ of a finite family of arcs Γ_1, the minimum distance $d(\Gamma_1(\varepsilon))$ increases as $\varepsilon\downarrow 0$ while the measure $|\Gamma_1(\varepsilon)|$ decreases. Moreover, the quantity $\rho(\Gamma_1(\varepsilon))$ does not change as ε varies and therefore, $\Gamma_1(\varepsilon)$ belongs to the class $\mathcal{F}$ for small enough ε's. Then, Theorem 5 contains Theorem 4.

Section 5 concludes the paper with some brief remarks on diverse questions related to Conjectures 1–3. In particular, the possible validity of n-dimensional versions of these conjectures is pointed out.

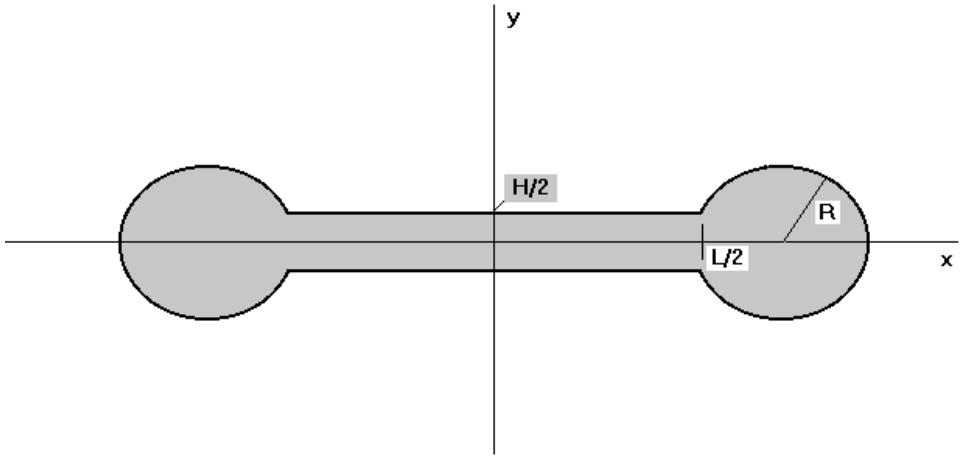


FIG. 1.

2. The role of the convexity of the domain. The convexity of the domain Ω is a condition which cannot be dropped if the property of the solutions to problem (1) stated by Conjecture 1 was generally true. To clarify this point, in this section an example is exhibited of a nonconvex domain where the conjecture fails. Thus, let us consider a domain Ω with the shape depicted in Figure 1. For this domain Ω, the solutions to problem (1) corresponding to different Γ_1's will be estimated by constructing appropriate sub- and supersolutions. By a supersolution to problem (1) we mean a function $w \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ satisfying

$$\left| \begin{array}{ll} \Delta w(x) \leq 0, & x \in \Omega, \\ w(x) \geq 0, & x \in \Gamma_0, \\ \frac{\partial w}{\partial n}(x) \geq 1, & x \in \Gamma_1. \end{array} \right. \tag{11}$$

A function $v \in \mathcal{C}^2(\Omega) \cap \mathcal{C}(\overline{\Omega})$ that satisfies the reverse inequalities is said to be a subsolution to problem (1). Let v and w be a sub and a supersolution, respectively, to problem (1). Since the domain Ω we are considering satisfies an interior sphere condition at every point of its boundary, the strong maximum principle and Hopf's lemma (see [1], [9], [13]) guarantee that the inequalities

$$v(x) \leq u(x) \leq w(x), \ x \in \overline{\Omega}$$

are satisfied by the solution u to problem (1). In the discussion below, affine, quadratic, and potential sub- and supersolutions are utilized which are simple enough to make the calculations feasible and, at the same time, to provide estimates which suffice for our purpose. As a general presentation of the method of sub- and supersolution to obtain approximations in boundary value problems of elliptic type, we cite [13]. For a more systematic study of estimates of the solution to problems like (1) using potential or affine sub- and supersolution we refer to [3].

As can be observed in Figure 1, the geometry of the domain under consideration is completely determined by the positive parameters L, R, and H $(R > H/2)$. Note that the smallest ball centered at the origin that contains Ω is given by $B_{r_0}(0)$ with $r_0 = L/2 + R + \sqrt{R^2 - H^2/4}$. This fact will be used later to obtain appropriate estimates. Moreover, we denote by a the length of Γ_1; i.e., $a = |\Gamma_1|$, assuming that

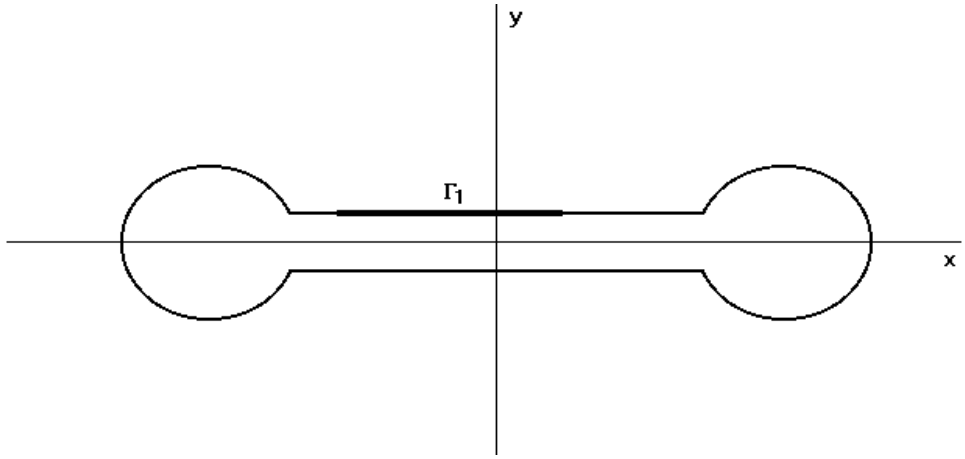


FIG. 2.

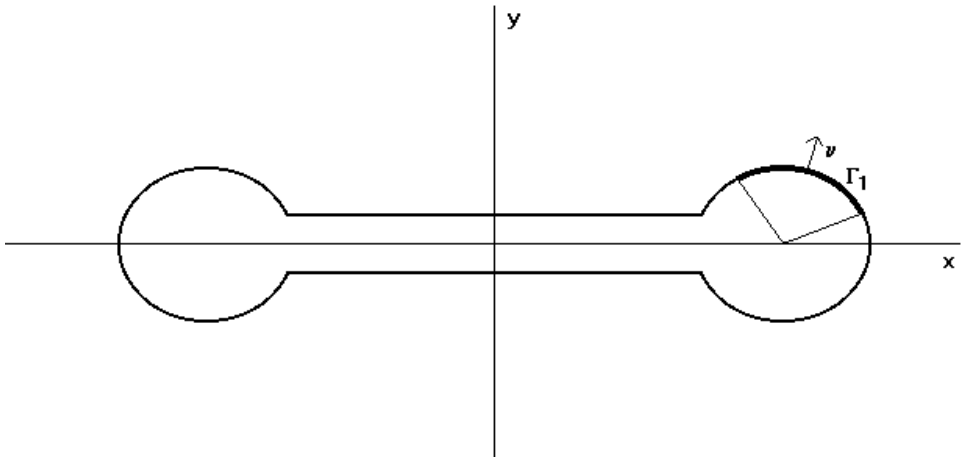


FIG. 3.

$a < \min\{L,\ \pi R/2\}$. First we will construct several supersolutions to problem (1) which depend on the relative position occupied by a connected arc Γ_1 on $\partial\Omega$. By the symmetry of the domain and the assumption made on a, it will be sufficient to analyze three situations which are respectively characterized by Figures 2, 3, and 4.

In the following discussion, we always denote by u the solution to the instance of problem (1) that is being analyzed.

Consider first the case illustrated in Figure 2. The function $v_1(x,y) = R + y$ is a supersolution to problem (1) in this case. In fact, v_1 is harmonic and

$$(v_1|_{\Gamma_0})(x,y) \geq R - R = 0,\ (x,y) \in \Gamma_0,$$
$$\left(\frac{\partial v_1}{\partial n}\right)\Big|_{\Gamma_1}(x,y) = 1,\ (x,y) \in \Gamma_1.$$

Thus, $u \leq v_1$ in $\overline{\Omega}$ and then

$$\sup_{\Omega} u \leq \sup_{\Omega} v_1 = 2R. \tag{12}$$

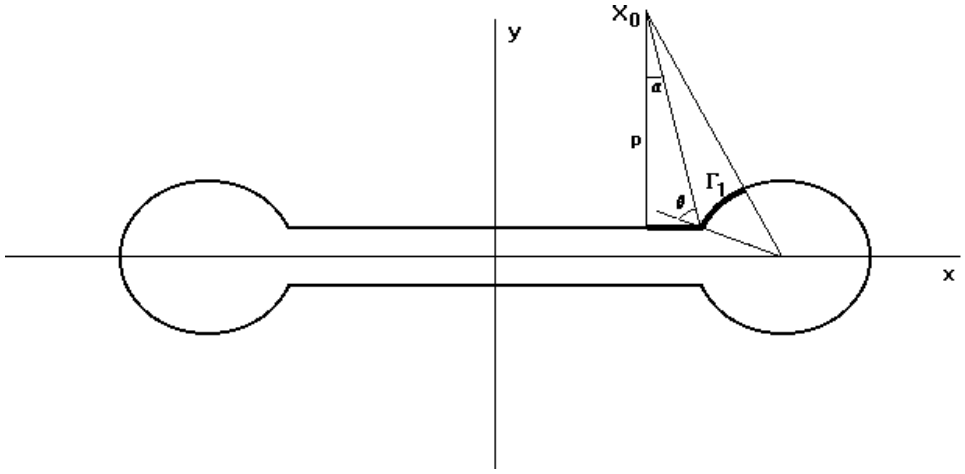


FIG. 4.

When Γ_1 is an arc of circle (see Figure 3) we denote by ν the unit outward-pointing normal vector to the circle corresponding to the midpoint of Γ_1. Then, the affine function

$$v_2(x,y) = \frac{1}{\cos(a/(2R))}\left(\frac{L}{2} + R + \sqrt{R^2 - \frac{H^2}{4}} + \langle \nu, (x,y)\rangle\right),$$

where $\langle\cdot,\cdot\rangle$ indicates the inner product in $\mathbb{R}^2$, is a supersolution to problem (1). Indeed, from the above-mentioned inclusion $\Omega \subset B_{r_0}(0)$, $r_0 = L/2 + R + \sqrt{R^2 - H^2/4}$ and using the Cauchy–Schwarz inequality we deduce

$$(v_2|_{\Gamma_0})(x,y) \geq \frac{1}{\cos(a/(2R))}\left(\frac{L}{2} + R + \sqrt{R^2 - \frac{H^2}{4}} - |(x,y)|\right) \geq 0,\ (x,y) \in \Gamma_0,$$

and, if $n(x,y)$ indicates the unitary outward-pointing normal vector at the point $(x,y) \in \partial\Omega$,

$$\left(\frac{\partial v_2}{\partial n}\right)\Big|_{\Gamma_1}(x,y) = \frac{\langle \nu, n(x,y)\rangle}{\cos(a/(2R))} \geq \frac{\cos(a/(2R))}{\cos(a/(2R))} = 1,\ (x,y) \in \Gamma_1.$$

Therefore,

$$\sup_{\Omega} u \leq \sup_{\Omega} v_2 \leq \frac{2\left(L/2 + R + \sqrt{R^2 - H^2/4}\right)}{\cos(a/(2R))} \leq \frac{L + 4R}{\cos(a/(2R))}. \tag{13}$$

The second inequality in (13) follows from a new application of the Cauchy–Schwarz inequality. For the case described by Figure 4 we propose a potential supersolution of the form (cf. [3])

$$v_3(x,y) = \frac{D(X_0,\Gamma_1)}{h(X_0,\Gamma_1)}\ \ln\left(\frac{D(X_0,\Gamma_0)}{|(x,y) - X_0|}\right),$$

where X_0 is a point not belonging to $\overline{\Omega}$ and, if Γ is a subset of $\partial\Omega$,

$$D(X_0,\Gamma)=\sup_{(x,y)\in\Gamma}|(x,y)-X_0|\,,\quad h(X_0,\Gamma)=\inf_{(x,y)\in\Gamma}\left\langle\frac{X_0-(x,y)}{|(x,y)-X_0|},n(x,y)\right\rangle.$$

In fact, we obviously have

$$v_3(x,y)\geq 0,\ (x,y)\in\Gamma_0,$$

and placing X_0 as in Figure 4 we obtain

$$\begin{aligned}\left(\frac{\partial v_3}{\partial n}\right)\Big|_{\Gamma_1}(x,y)&=\frac{D(X_0,\Gamma_1)}{h(X_0,\Gamma_1)}\,\frac{1}{|(x,y)-X_0|}\left\langle\frac{X_0-(x,y)}{|(x,y)-X_0|},n(x,y)\right\rangle\\&\geq\frac{D(X_0,\Gamma_1)}{|(x,y)-X_0|}\geq 1,\qquad (x,y)\in\Gamma_1.\end{aligned}$$

Thus, the following estimate for u is deduced:

$$\sup_\Omega u\leq\sup_\Omega v_3\leq\frac{D(X_0,\Gamma_1)}{h(X_0,\Gamma_1)}\,\ln\left(\frac{D(X_0,\Gamma_0)}{d(X_0,\Omega)}\right),\tag{14}$$

where, as usual, $d(X_0,\Omega)$ denotes the distance of X_0 to Ω. Now, simple calculations show that an upper bound independent of the parameter H can be found for the estimate of $\sup_\Omega u$ given by (14). This property, which is also shared by the bounds provided by (12) and (13), is to be exploited further on to complete the argument. To find such an upper bound we set $a=a_1+a_2$ with a_2 denoting the length of the part of Γ_1 placed on the circle. Let $(L/2-a_1,H/2+p)$ be the coordinates of X_0 with the value of p given by

$$p=\left(a_1+\sqrt{R^2-\frac{H^2}{4}}\right)\tan\left(\frac{a_2}{R}+\arcsin\left(\frac{H}{2R}\right)\right)-\frac{H}{2};$$

then we can write

$$D(X_0,\Gamma_1)=\sqrt{a_1^2+p^2},\tag{15}$$

$$D(X_0,\Gamma_0)=R+\sqrt{\left(p+\tfrac{H}{2}\right)^2+\left(L-a_1+\sqrt{R^2-\tfrac{H^2}{4}}\right)^2},\tag{16}$$

$$d(X_0,\Omega)=\min\left\{p,\sqrt{\left(a_1+\sqrt{R^2-\tfrac{H^2}{4}}\right)^2+\left(p+\tfrac{H}{2}\right)^2}-R\right\}.\tag{17}$$

Furthermore, taking into account that

$$\cos\alpha=\tfrac{p}{\sqrt{a_1^2+p^2}},\quad\cos\theta=\tfrac{H}{2R}\,\cos\alpha+\sqrt{1-\tfrac{H^2}{4R^2}}\,\sin\alpha\geq\sqrt{1-\tfrac{H^2}{4R^2}}\,\tfrac{a_1}{\sqrt{a_1^2+p^2}},$$

where α and θ are the angles specified in Figure 4, we obtain

$$h(X_0,\Gamma_1)=\min\{\cos\alpha,\cos\theta\}\geq\frac{1}{\sqrt{a_1^2+p^2}}\min\left\{p,a_1\sqrt{1-\frac{H^2}{4R^2}}\right\}.\tag{18}$$

By introducing (15)–(18) in (14) and taking $H < R$, after some algebraic manipulations we conclude

$$\sup_{\Omega} u \leq \frac{a_1^2 + p^2}{\min\{p, \frac{\sqrt{3}}{2}a_1\}} \ln\left(\frac{\frac{5}{2}R + p + L - a_1}{\min\left\{p, \sqrt{\left(a_1 + \frac{\sqrt{3}}{2}R\right)^2 + p^2} - R\right\}} \right). \tag{19}$$

Finally, in view of $\lim_{H\downarrow 0} p = (a_1 + R)\tan(a_2/R)$, an estimate independent of H for $\sup_{\Omega} u$ can be easily deduced from (19).

Next, we will consider the instance of problem (1) in which Γ_1 splits into two symmetric segments of length $a/2$ as shown in Figure 5. On the subdomain $\Omega' = (-a/4, a/4) \times (-H/2, H/2)$ (in gray in Figure 5), the harmonic function

$$w(x,y) = \frac{1}{H}\left(y^2 - x^2 + \frac{a^2}{16} - \frac{H^2}{4}\right)$$

satisfies

$$\left(\frac{\partial w}{\partial n}\right)\Big|_{\Gamma_1}(x,y) = 1,\ (x,y) \in \Gamma_1,$$
$$w\left(\pm\frac{a}{4}, y\right) = \frac{1}{H}\left(y^2 - \frac{H^2}{4}\right) \leq 0,\ |y| < \frac{H}{2}.$$

By the strong maximum principle and Hopf's lemma (see [13], [9], [1]), the solution u to problem (1) satisfies $u > 0$ in Ω; hence

$$u\left(\pm\frac{a}{4}, y\right) > 0,\ |y| < \frac{H}{2};$$

moreover, the symmetry of Ω and Γ_1 guarantees that $\sup_{\Omega} u = u(0, \pm H/2)$. Then we see that w is a subsolution in Ω' of u and therefore,

$$\sup_{\Omega} u \geq \sup_{\Omega'} w = w\left(0, \pm\frac{H}{2}\right) = \frac{a^2}{16H}. \tag{20}$$

Now, the fact that a domain shaped like that in Figure 1 does not satisfy the property stated in Conjecture 1 quickly follows from inequalities (12), (13), (19), and (20). In effect, the lower bound for $\sup_{\Omega} u$ given by (20) depends on the reciprocal of H, unlike the upper bounds provided by the remaining inequalities. Therefore, Conjecture 1 is violated by the domain Ω under consideration when a sufficiently small H is chosen.

3. Proof of Theorem 4. The technique we employ to prove Theorem 4 consists of several steps. Broadly speaking, we use Ghizzetti's exact solution to problem (1) in the circle and conformal maps to estimate the right-hand side of (7) for small enough ε. An estimate of the other side is obtained through the maximum principles and more involved estimations of normal derivatives. To begin with, using the notation settled in the Introduction, for $k = 1, 2, \ldots, n$ let us consider the mixed boundary value problems

$$\left|\begin{array}{ll} \Delta u_k(x) = 0, & x \in \Omega, \\ u_k(x) = 0, & x \in \Gamma_0, \\ \frac{\partial u_k}{\partial n}(x) = 1, & x \in \Gamma_1^{(k)}, \\ \frac{\partial u_k}{\partial n}(x) = 0, & x \in \Gamma_1^{(j)},\ j \neq k, \end{array}\right. \tag{21}$$

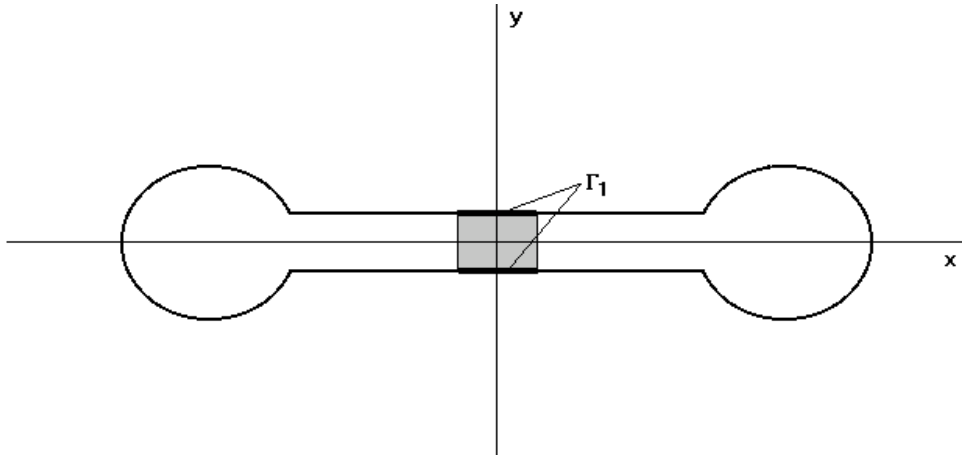


FIG. 5.

and

$$
(22)\qquad \left|\begin{array}{ll} \Delta v_k(x)=0, & x\in\Omega,\\ v_k(x)=0, & x\in\Gamma_0,\\ \frac{\partial v_k}{\partial n}(x)=1, & x\in\Gamma_1^{(k)},\\ v_k(x)=0, & x\in\Gamma_1^{(j)},\ j\neq k. \end{array}\right.
$$

The normal derivative $\partial v_k/\partial n$ of the solution v_k to (22) is a bounded and (at least) continuous function on Γ_1 and therefore, it makes sense to consider also the following problem:

$$
(23)\qquad \left|\begin{array}{ll} \Delta w_k(x)=0, & x\in\Omega,\\ w_k(x)=0, & x\in\Gamma_0,\\ \frac{\partial w_k}{\partial n}(x)=0, & x\in\Gamma_1^{(k)},\\ \frac{\partial w_k}{\partial n}(x)=-\frac{\partial v_k}{\partial n}(x), & x\in\Gamma_1^{(j)},\ j\neq k. \end{array}\right.
$$

From the strong maximum principle and Hopf's lemma (see [9],[13]), we deduce that the functions $u_k,\ v_k$, and w_k are nonnegative on $\overline{\Omega}$. Take, for instance, the function u_k. In view of the imposed boundary conditions, u_k is not constant on Ω; then, $\inf_\Omega u_k = u_k(x_0)$ for a certain $x_0\in\partial\Omega$ by the strong maximum principle. But Hopf's lemma ensures the normal derivative $(\partial u_k/\partial n)(x_0)$ is negative at the point x_0, so that $x_0\in\Gamma_0$ and therefore $\inf_\Omega u_k = u_k(x_0)=0$. The nonnegativity of v_k and w_k follows in a similar way.

Several relationships hold among functions $u_k,\ v_k, w_k$, and the solution $u[\Gamma_1]$ to (1). The most useful of these are listed in the following lemma.

LEMMA 6. *If $u[\Gamma_1],\ u_k,\ v_k$, and w_k respectively denote the solutions to problems* (1), (21), (22), *and* (23), *then the following relationships hold:*
(i) $u[\Gamma_1]=\sum_{k=1}^n u_k$;
(ii) $u_k = v_k + w_k,\ k=1,2,\dots,n$;
(iii) $v_k \le u_k \le v_k + \max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_\infty \sum_{j\neq k} u_j$.

Furthermore, if the inequality $\sum_{k=1}^n \max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_\infty < 1$ *holds, then*

(iv)

$$\max_{1\le k\le n}\sup_{\Omega} v_k \le \sup_{\Omega} u[\Gamma_1] \le \frac{\max_{1\le k\le n}\sup_{\Omega} v_k}{1-\sum_{k=1}^{n}\max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_{\infty}}.$$

Proof. Assertions (i) and (ii) are immediate from the definitions of $u[\Gamma_1]$, u_k, v_k, and w_k. The first inequality in (iii) follows from (ii) and the nonnegativity of w_k. To derive the second inequality in (iii), we consider the function $\widetilde{w}_k$ that solves the problem

$$\left|\begin{array}{ll} \Delta\widetilde{w}_k(x)=0, & x\in\Omega,\\ \widetilde{w}_k(x)=0, & x\in\Gamma_0,\\ \frac{\partial\widetilde{w}_k}{\partial n}(x)=0, & x\in\Gamma_1^{(k)},\\ \frac{\partial w_k}{\partial n}(x)=\max_{j\neq k}\left\|\frac{\partial v_k}{\partial n}\big|_{\Gamma_1^{(j)}}\right\|_{\infty}, & x\in\Gamma_1^{(j)},\ j\neq k. \end{array}\right.$$

The strong maximum principle and Hopf's lemma provide $w_k\le\widetilde{w}_k$ on $\overline{\Omega}$. On the other hand, it is obvious that

$$\widetilde{w}_k=\max_{j\neq k}\left\|\frac{\partial v_k}{\partial n}\Big|_{\Gamma_1^{(j)}}\right\|_{\infty}\sum_{j\neq k}u_j,$$

and thus, from (ii) we obtain

$$u_k=v_k+w_k\le v_k+\widetilde{w}_k=v_k+\max_{j\neq k}\left\|\frac{\partial v_k}{\partial n}\Big|_{\Gamma_1^{(j)}}\right\|_{\infty}\sum_{j\neq k}u_j.$$

In order to prove the inequalities (iv), we take $\sum_{k=1}^{n}$ in the second inequality (iii) to obtain

$$\sum_{k=1}^{n}u_k\le\sum_{k=1}^{n}v_k+\sum_{k=1}^{n}\left(\max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_{\infty}\sum_{j\neq k}u_j\right), \tag{24}$$

or, taking into account (i) and the nonnegativity of u_k, $k=1,2,\dots,n$,

$$\begin{aligned} u[\Gamma_1] &\le \sum_{k=1}^{n}v_k+\sum_{k=1}^{n}\left(\max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_{\infty}\sum_{j\neq k}u_k\right)\\ &\le \sum_{k=1}^{n}v_k+u[\Gamma_1]\sum_{k=1}^{n}\max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_{\infty}. \end{aligned} \tag{25}$$

By the strong maximum principle and Hopf's lemma, there exists a point $P\in\Gamma_1$ such that $\sup_{\Omega}u[\Gamma_1]=u[\Gamma_1](P)$. Without loss of generality we can assume $P\in\Gamma_1^{(1)}$. From (25) we then obtain

$$\sup_{\Omega}u[\Gamma_1]=u[\Gamma_1](P)\le v_1(P)+u[\Gamma_1](P)\sum_{k=1}^{n}\max_{j\neq k}\left\|(\partial v_k/\partial n)|_{\Gamma_1^{(j)}}\right\|_{\infty},$$

and using the hypothesis $\sum_{k=1}^{n} \max_{j\neq k} \left\| (\partial v_k/\partial n)|_{\Gamma_1^{(j)}} \right\|_\infty < 1$,

$$\begin{aligned}\sup_\Omega u[\Gamma_1] &\leq \frac{v_1(P)}{1-\sum_{k=1}^{n} \max_{j\neq k} \left\| (\partial v_k/\partial n)|_{\Gamma_1^{(j)}} \right\|_\infty} \\ &\leq \frac{\sup_\Omega v_1}{1-\sum_{k=1}^{n} \max_{j\neq k} \left\| (\partial v_k/\partial n)|_{\Gamma_1^{(j)}} \right\|_\infty} \\ &\leq \frac{\max_{1\leq k\leq n} \sup_\Omega v_k}{1-\sum_{k=1}^{n} \max_{j\neq k} \left\| (\partial v_k/\partial n)|_{\Gamma_1^{(j)}} \right\|_\infty},\end{aligned}$$

which proves the second inequality in (iv). As for the first inequality in (iv), from the first one in (iii) we deduce

$$\sup_\Omega v_k \leq \sup_\Omega u[\Gamma_1],$$

whence

$$\max_{1\leq k\leq n} \sup_\Omega v_k \leq \sup_\Omega u[\Gamma_1].$$

This completes the proof. □

The second inequality from Lemma 6(iv) can be employed to prove Theorem 4. In fact, assume for a moment that the expressions

$$\sup_\Omega u[\Gamma_1^*(\varepsilon)] = C_0\ ((b-a)\,\varepsilon)^p + o(\varepsilon^p), \tag{26}$$

$$\left\| (\partial v_k/\partial n)|_{\Gamma_1^{(j)}(\varepsilon)} \right\|_\infty \leq C_{j.k}\ ((b_k-a_k)\,\varepsilon)^q + o(\varepsilon^q) \tag{27}$$

hold for sufficiently small ε's with $p,\ q > 0$ and constants C_0 and $C_{j.k}$ which are independent of a, b and a_k, b_k, $k = 1, 2, \ldots, n$, respectively. A proof of (26) and (27) for a Dini-smooth domain (with $p = 1$ and $q = 2$) is furnished by Theorem 8 below. With the inequalities (26) and (27) at hand, we succeed in proving Theorem 4 as follows.

Proof of Theorem 4. By using (26), $\sup_\Omega v_k$ admits the following representation:

$$\sup_\Omega v_k = C_0\ ((b_k-a_k)\,\varepsilon)^p + o(\varepsilon^p),\ k = 1, 2, \ldots, n,$$

and, since the hypothesis of item (iv) of Lemma 6 is satisfied when ε is small enough (for instance, $\varepsilon^q < 1/[2\sum_{k=1}^{n}(b_k-a_k)^q \max_{j\neq k} C_{j.k}]$), the second inequality from Lemma 6(iv) and (27) provide for these ε's

$$\begin{aligned}\sup_\Omega u[\Gamma_1(\varepsilon)] &\leq \frac{C_0 \max_{1\leq k\leq n} ((b_k-a_k)\,\varepsilon)^p + o(\varepsilon^p)}{1-\sum_{k=1}^{n} \max_{j\neq k} C_{j.k}\ ((b_k-a_k)\,\varepsilon)^q + o(\varepsilon^q)} \\ &= C_0 \max_{1\leq k\leq n} ((b_k-a_k)\,\varepsilon)^p + o(\varepsilon^p). \qquad (28)\end{aligned}$$

By recalling that $b-a = \sum_{k=1}^{n}(b_k-a_k)$ and $b_k-a_k > 0,\ k = 1, 2, \ldots, n$, from (26) and (28) we get

$$\begin{aligned}\sup_\Omega u[\Gamma_1(\varepsilon)] &\leq C_0 \max_{1\leq k\leq n} ((b_k-a_k)\,\varepsilon)^p + o(\varepsilon^p) \\ &< C_0\ ((b-a)\,\varepsilon)^p + o(\varepsilon^p) = \sup_\Omega u[\Gamma_1^*(\varepsilon)],\end{aligned}$$

for sufficiently small ε's, as Theorem 4 asserts. □

The remainder of this section is devoted to justifying the validity of the expressions (26) and (27) under reasonable hypotheses on the regularity of the domain Ω. Let us begin by showing that (26) and (27) hold for $\Omega = B_1(0)$, the circle of radius 1; then, the technique of conformal maps will be applied to extend their validity to Dini-smooth domains which satisfy an interior sphere condition. Thus, we assume that $\Omega = B_1(0)$ in the following. As usual, polar coordinates are employed to denote points belonging to the circle.

An explicit solution to the general mixed boundary value problem

$$\left|\begin{array}{ll} \Delta u(\rho,\phi) = \frac{1}{\rho}\left((\rho u_\rho)_\rho + \frac{1}{\rho}\, u_{\phi\phi}\right) = 0, & (\rho,\phi) \in B_1(0), \\ u(1,\phi) = F(\phi), & \alpha < \phi < 2\pi, \\ \frac{\partial u}{\partial \rho}(1,\phi) = G(\phi), & 0 < \phi < \alpha \end{array}\right. \tag{29}$$

was found by Ghizzetti in [7] (see also [8]). Under suitable hypotheses of regularity on the functions F and G it is shown by this author that the solution $u(\rho,\phi)$ to (29) can be expressed as follows:

$$\begin{aligned} u(\rho,\phi) = {} & \tfrac{1}{2\pi}\int_0^\alpha \ln\left(\frac{\sqrt{H(\rho,\phi,\theta)}+\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}{\sqrt{H(\rho,\phi,\theta)}-\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}\right) G(\theta)\, d\theta \\ & + \tfrac{1}{2\pi}\,\frac{1-\rho^2}{\sqrt{M(\rho,\phi)+N(\rho,\phi)}}\int_\alpha^{2\pi} \frac{\sqrt{H(\rho,\phi,\theta)}+\sqrt{K(\rho,\phi,\theta)}}{1-2\rho\cos(\phi-\theta)+\rho^2}\, F(\theta)\, d\theta, \end{aligned} \tag{30}$$

where

$$H(\rho,\phi,\theta) = \frac{\sin(|\alpha-\theta|/2)}{\sin(\theta/2)}\,(1-2\rho\cos\phi+\rho^2), \tag{31}$$

$$K(\rho,\phi,\theta) = \frac{\sin(\theta/2)}{\sin(|\alpha-\theta|/2)}\,(1-2\rho\cos(\phi-\alpha)+\rho^2), \tag{32}$$

$$M(\rho,\phi) = 2\sqrt{(1-2\rho\cos\phi+\rho^2)(1-2\rho\cos(\phi-\alpha)+\rho^2)}, \tag{33}$$

$$N(\rho,\phi) = 2\,(\cos(\alpha/2) - 2\rho\cos(\phi-\alpha/2)+\rho^2\cos(\alpha/2)). \tag{34}$$

Note that $M = 2\sqrt{HK}$. From (30) and (31)–(34) we easily derive an explicit solution to the instance of problem (1) in which $\Omega = B_1(0)$ and $\Gamma_1 = (0,\alpha)$. In fact, denoting this solution by u_α we can write

$$u_\alpha(\rho,\phi) = \tfrac{1}{2\pi}\int_0^\alpha \ln\left(\frac{\sqrt{H(\rho,\phi,\theta)}+\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}{\sqrt{H(\rho,\phi,\theta)}-\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}\right) d\theta, \tag{35}$$

where H, K, M, and N are defined by (31)–(34). Now we prove that the supremum $\sup_{B_1(0)} u_\alpha$ is attained at the boundary point $(1,\alpha/2)$. In fact, the strong maximum principle and Hopf's lemma show that $\sup_{B_1(0)} u_\alpha$ is attained at a point $(1,\phi_0)$ such that $0 < \phi_0 < \alpha$, and so, we must look for the maximum of the boundary value $\phi \mapsto h(\phi) = u_\alpha(1,\phi)$, $0 < \phi < \alpha$. Since u_α is symmetric with respect to the line $\phi = \alpha/2$, the point $\phi = \alpha/2$ is a point of symmetry for the function h; i.e., $h(\alpha/2-\phi) = h(\alpha/2+\phi)$, $0 \le \phi < \alpha/2$. Furthermore, h is a strictly concave function in $(0,\alpha)$ which, jointly with its symmetry, implies that $\sup_{B_1(0)} u_\alpha = \sup_{0<\phi<\alpha} h(\phi) = h(\alpha/2) = u_\alpha(1,\alpha/2)$. That h is really a strictly concave function can be seen by using once again the maximum principle and Hopf's lemma. On the one hand, h is a smooth function and then the equation

$$\Delta u_\alpha(\rho,\phi) = \frac{1}{\rho}\left(\left(\rho\,(u_\alpha)_\rho\right)_\rho + \frac{1}{\rho}\,(u_\alpha)_{\phi\phi}\right) = 0$$

holds up to the boundary $\rho = 1,\ 0 < \phi < \alpha$, so that we can write

$$h''(\phi) = (u_\alpha)_{\phi\phi}(1,\phi) = -\left(\rho\,(u_\alpha)_\rho\right)_\rho(1,\phi),\ 0 < \phi < \alpha. \tag{36}$$

On the other hand, the function defined by $v = \rho\,(u_\alpha)_\rho,\ (\rho,\phi) \in B_1(0)$, is harmonic and its maximum value (equal to 1) on $\overline{B_1(0)}$ is reached at every point $(1,\phi)$, $0 < \phi < \alpha$. Thus, Hopf's lemma shows that the normal derivative $(v)_\rho(1,\phi) = \left(\rho\,(u_\alpha)_\rho\right)_\rho(1,\phi) > 0$, $0 < \phi < \alpha$, and so, from (36) we obtain $h''(\phi) < 0,\ 0 < \phi < \alpha$; that is, h is strictly concave.

Once it is known that $\sup_{B_1(0)} u_\alpha = u_\alpha(1,\alpha/2)$, Ghizzetti's formula (36) becomes useful in deriving an explicit expression for $\sup_{B_1(0)} u_\alpha$. Indeed, we have

$$\begin{aligned}
&\sup_{B_1(0)} u_\alpha = u_\alpha(1,\alpha/2)\\
&= \tfrac{1}{2\pi}\int_0^\alpha \ln\left(\frac{\sqrt{H(1,\alpha/2,\theta)}+\sqrt{M(1,\alpha/2)-N(1,\alpha/2)}+\sqrt{K(1,\alpha/2,\theta)}}{\sqrt{H(1,\alpha/2,\theta)}-\sqrt{M(1,\alpha/2)-N(1,\alpha/2)}+\sqrt{K(1,\alpha/2,\theta)}}\right)d\theta,
\end{aligned}$$

or, realizing that

$$\begin{aligned}
H(1,\alpha/2,\theta) &= 2\,\tfrac{\sin(|\alpha-\theta|/2)}{\sin(\theta/2)}\,(1-\cos(\alpha/2)),\\
K(1,\alpha/2,\theta) &= 2\,\tfrac{\sin(\theta/2)}{\sin(|\alpha-\theta|/2)}\,(1-\cos(\alpha/2)),\\
M(1,\alpha/2) &= 4\,(1-\cos(\alpha/2)),\\
N(1,\alpha/2) &= 4\,(\cos(\alpha/2)-1) = -M(1,\alpha/2),
\end{aligned}$$

and after some simplifications,

$$\begin{aligned}
\sup_{B_1(0)} u_\alpha &= \tfrac{1}{\pi}\int_0^\alpha \ln\left|\frac{\sqrt{\sin((\alpha-\theta)/2)}+\sqrt{\sin(\theta/2)}}{\sqrt{\sin((\alpha-\theta)/2)}-\sqrt{\sin(\theta/2)}}\right|d\theta\\
&= \tfrac{\alpha}{\pi}\int_0^1 \ln\left|\frac{\sqrt{\sin(\frac{\alpha}{2}(1-\lambda))}+\sqrt{\sin(\frac{\alpha}{2}\lambda)}}{\sqrt{\sin(\frac{\alpha}{2}(1-\lambda))}-\sqrt{\sin(\frac{\alpha}{2}\lambda)}}\right|d\lambda.
\end{aligned} \tag{37}$$

The expression (35) may be further used to compute the radial derivative $\partial u_\alpha/\partial\rho$ at a point $(1,\phi)$ belonging to $\Gamma_0 = (\alpha, 2\pi)$. With this purpose, we note that $u_\alpha(1,\phi) = 0$ for every $\alpha < \phi < 2\pi$, and so, from (35) we obtain

$$\begin{aligned}
&\tfrac{\partial u_\alpha}{\partial\rho}(R,\phi) = \lim_{\rho\uparrow 1}\tfrac{u_\alpha(\rho,\phi)}{\rho-1}\\
&= \tfrac{1}{2\pi}\lim_{\rho\uparrow 1}\int_0^\alpha \ln\left(\frac{\sqrt{H(\rho,\phi,\theta)}+\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}{\sqrt{H(\rho,\phi,\theta)}-\sqrt{M(\rho,\phi)-N_R(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}\right)^{\frac{1}{\rho-1}}d\theta\\
&= \tfrac{1}{2\pi}\lim_{\rho\uparrow 1}\int_0^\alpha \ln\left(\frac{1+\frac{\sqrt{M(\rho,\phi)-N(\rho,\phi)}}{\sqrt{H(\rho,\phi,\theta)}+\sqrt{K(\rho,\phi)}}}{1-\frac{\sqrt{M(\rho,\phi)-N(\rho,\phi)}}{\sqrt{H(\rho,\phi,\theta)}+\sqrt{K(\rho,\phi)}}}\right)^{\frac{1}{\rho-1}}d\theta.
\end{aligned} \tag{38}$$

Routine calculations applied to the expressions (31)–(34) show that

$$\sqrt{M(\rho,\phi)-N(\rho,\phi)} = \frac{\sin(\alpha/2)}{\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}}\,(1-\rho)+O((1-\rho)^2)$$

and

$$\sqrt{H(\rho,\phi,\theta)}+\sqrt{K(\rho,\phi,\theta)}=2\left(\sin\tfrac{\phi}{2}\sqrt{\tfrac{\sin((\alpha-\theta)/2)}{\sin(\theta/2)}}+\sin\tfrac{\phi-\alpha}{2}\sqrt{\tfrac{\sin(\theta/2)}{\sin((\alpha-\theta)/2)}}\right)$$
$$+O(1-\rho),$$

whence

$$\begin{aligned}&\lim_{\rho\uparrow 1}\ln\left(\frac{1+\frac{\sqrt{M(\rho,\phi)-N(\rho,\phi)}}{\sqrt{H(\rho,\phi,\theta)+\sqrt{K(\rho,\phi)}}}}{1-\frac{\sqrt{M(\rho,\phi)-N(\rho,\phi)}}{\sqrt{H(\rho,\phi,\theta)+\sqrt{K(\rho,\phi)}}}}\right)^{\frac{1}{\rho-1}}\\&=-\frac{\sin(\alpha/2)}{\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}\left(\sin\frac{\phi}{2}\sqrt{\frac{\sin((\alpha-\theta)/2)}{\sin(\theta/2)}}+\sin\frac{\phi-\alpha}{2}\sqrt{\frac{\sin(\theta/2)}{\sin((\alpha-\theta)/2)}}\right)}.\end{aligned}\tag{39}$$

On the other hand, using the Taylor's series of $x\mapsto\ln((1+x)/(1-x))$ and the inequality $Ax+Bx^{-1}\geq 2\sqrt{AB}$ which holds for $A,B,x>0$, we see that for ρ close enough to 1 there exists a constant $L>0$ such that

$$\begin{aligned}&\left|\ln\left(\frac{1+\frac{\sqrt{M(\rho,\phi)-N(\rho,\phi)}}{\sqrt{H(\rho,\phi,\theta)+\sqrt{K(\rho,\phi)}}}}{1-\frac{\sqrt{M(\rho,\phi)-N(\rho,\phi)}}{\sqrt{H(\rho,\phi,\theta)+\sqrt{K(\rho,\phi)}}}}\right)^{\frac{1}{\rho-1}}\right|\\&\leq\frac{L\sin(\alpha/2)}{\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}\left(\sin\frac{\phi}{2}\sqrt{\frac{\sin((\alpha-\theta)/2)}{\sin(\theta/2)}}+\sin\frac{\phi-\alpha}{2}\sqrt{\frac{\sin(\theta/2)}{\sin((\alpha-\theta)/2)}}\right)}\\&\leq\frac{L\sin(\alpha/2)}{2\sin(\phi/2)\sin((\phi-\alpha)/2)},\ 0<\theta<\alpha.\end{aligned}\tag{40}$$

In view of (39) and (40) we can apply the dominated convergence theorem to the last member of (38) to obtain

$$\begin{aligned}\frac{\partial u_\alpha}{\partial\rho}(1,\phi)&=-\frac{\sin(\alpha/2)}{2\pi\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}}\int_0^\alpha\frac{d\theta}{\sin\frac{\phi}{2}\sqrt{\frac{\sin((\alpha-\theta)/2)}{\sin(\theta/2)}}+\sin\frac{\phi-\alpha}{2}\sqrt{\frac{\sin(\theta/2)}{\sin((\alpha-\theta)/2)}}}\\&=-\frac{\alpha\sin(\alpha/2)}{2\pi\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}}\int_0^1\frac{d\lambda}{\sin\frac{\phi}{2}\sqrt{\frac{\sin(\alpha(1-\lambda)/2)}{\sin(\alpha\lambda/2)}}+\sin\frac{\phi-\alpha}{2}\sqrt{\frac{\sin(\alpha\lambda/2)}{\sin(\alpha(1-\lambda)/2)}}}.\end{aligned}\tag{41}$$

Now we are in the position to prove that expressions like (26) and (27) hold for the circle.

THEOREM 7. *By employing the above notation, we have*

$$\sup_{B_1(0)}u_\alpha=\tfrac{1}{2}\alpha+o(\alpha).\tag{42}$$

Moreover, if $\alpha<\phi_1<\phi_2<2\pi$, *then*

$$\sup_{\phi_1<\phi<\phi_2}\left|\frac{\partial u_\alpha}{\partial\rho}(1,\phi)\right|\leq\frac{1}{8\pi\min_{\phi\in[\phi_1,\phi_2]}[\sin(\phi/2)\sin((\phi-\alpha)/2)]}\,\alpha^2+o(\alpha^2).\tag{43}$$

Proof. With the purpose of proving the validity of expression (42), we observe that

$$\lim_{\alpha\downarrow 0}\int_0^1\ln\left|\frac{\sqrt{\sin(\frac{\alpha}{2}(1-\lambda))}+\sqrt{\sin(\frac{\alpha}{2}\lambda)}}{\sqrt{\sin(\frac{\alpha}{2}(1-\lambda))}-\sqrt{\sin(\frac{\alpha}{2}\lambda)}}\right|d\lambda=\int_0^1\ln\left|\frac{\sqrt{1-\lambda}+\sqrt{\lambda}}{\sqrt{1-\lambda}-\sqrt{\lambda}}\right|d\lambda=\tfrac{\pi}{2}.$$

The expression (42) immediately follows from this equality and from (37). On the other hand, it follows from the arithmetic mean-geometric mean inequality that, for every $0 < \lambda < 1$ and $\alpha < \phi < 2\pi$,

$$\sin\tfrac{\phi}{2}\sqrt{\tfrac{\sin(\alpha(1-\lambda)/2)}{\sin(\alpha\lambda/2)}} + \sin\tfrac{\phi-\alpha}{2}\sqrt{\tfrac{\sin(\alpha\lambda/2)}{\sin(\alpha(1-\lambda)/2)}} \geq 2\sqrt{\sin\tfrac{\phi}{2}\sin\tfrac{\phi-\alpha}{2}}\ ,$$

and hence

$$\int_0^1 \frac{d\lambda}{\sin\frac{\phi}{2}\sqrt{\frac{\sin(\alpha(1-\lambda)/2)}{\sin(\alpha\lambda/2)}}+\sin\frac{\phi-\alpha}{2}\sqrt{\frac{\sin(\alpha\lambda/2)}{\sin(\alpha(1-\lambda)/2)}}} \leq \frac{1}{2\sqrt{\sin\frac{\phi}{2}\sin\frac{\phi-\alpha}{2}}}. \tag{44}$$

Therefore, from (41) and (44) we see that

$$\begin{aligned}\left|\frac{\partial u_\alpha}{\partial\rho}(1,\phi)\right| &\leq \frac{\alpha\sin(\alpha/2)}{2\pi\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}}\,\frac{1}{2\sqrt{\sin(\phi/2)\sin((\phi-\alpha)/2)}}\\ &= \frac{1}{8\pi\sin(\phi/2)\sin((\phi-\alpha)/2)}\,\alpha^2+o(\alpha^2),\end{aligned}$$

whence we finally get expression (43). □

It should be observed that if $u_{R,\alpha}$ is the solution to problem (1) for $\Omega = B_R(0)$ and $\Gamma_1 = (0, R\alpha)$, then we have $u_{R,\alpha}(\rho,\phi) \equiv R\,u_\alpha(\rho/R,\phi)$. In this way, by setting $\varepsilon = |\Gamma_1| = R\alpha$, from Theorem 7 we conclude

$$\sup_{B_R(0)} u_{R,\alpha} = \tfrac{1}{2}\varepsilon + o(\varepsilon), \tag{45}$$

and, for $\alpha < \phi_1 < \phi_2 < 2\pi$,

$$\sup_{\phi_1<\phi<\phi_2}\left|\frac{\partial u_{R,\alpha}}{\partial\rho}(R,\phi)\right| \leq \frac{1}{8\pi R^2\min_{\phi\in[\phi_1,\phi_2]}[\sin(\phi/2)\sin((\phi-\alpha)/2)]}\,\varepsilon^2+o(\varepsilon^2). \tag{46}$$

Now we turn to consider the case of a Jordan domain Ω with a Dini-smooth boundary curve γ. Choose a point $O \in \gamma$ and take O as the origin of arcs. To facilitate the following calculations $\gamma(s)$ is given in complex form, $\gamma(s) = \gamma_1(s) + i\gamma_2(s)$. By translating Ω if necessary, we can take O as the origin of coordinates; i.e., we set $O = \gamma'(0) = (0,0)$. Furthermore, a rotation of Ω around O can be done in such a way so as to get $\gamma'(0) = 1$. As before, we symbolize by $u[\Gamma_1]$ the solution to problem (1) for $\Gamma_1 = (0,\varepsilon)$. Among the conformal maps f from $B_1(0)$ onto Ω, we select that one which verifies $f(1) = 0$ and $f'(1) = -i\,|f'(1)|$. In this way, the function $v = u[\Gamma_1]\circ f$ satisfies

$$\left|\begin{array}{cc}\Delta v(\rho,\phi) = 0, & (\rho,\phi)\in B_1(0),\\ v(1,\phi) = 0, & \alpha(\varepsilon)<\phi<2\pi,\\ \frac{\partial v}{\partial\rho}(1,\phi) = \left|f'(e^{i\phi})\right|, & 0<\phi<\alpha(\varepsilon),\end{array}\right. \tag{47}$$

where $\alpha(s)$ is such that $f(e^{i\alpha(s)}) = \gamma(s)$, $0 \leq s \leq |\partial\Omega|$. Differentiating this last equality with respect to s, we obtain $f'(e^{i\alpha(s)})\,e^{i\alpha(s)}i\alpha'(s) = \gamma'(s)$ and then, from the assumptions made on γ and f, we deduce $\alpha(0) = 0$ and $\alpha'(0) = 1/|f'(1)|$, so that for $\varepsilon \downarrow 0$ we can write

$$\alpha(\varepsilon) = \frac{1}{|f'(1)|}\varepsilon + o(\varepsilon). \tag{48}$$

By Ghizzetti's formula (30), the function v solving (47) can be represented as

$$v(\rho,\phi) = \tfrac{1}{2\pi}\int_0^{\alpha(\varepsilon)} \ln\left(\frac{\sqrt{H(\rho,\phi,\theta)}+\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}{\sqrt{H(\rho,\phi,\theta)}-\sqrt{M(\rho,\phi)-N(\rho,\phi)}+\sqrt{K(\rho,\phi,\theta)}}\right)\left|f'(e^{i\theta})\right|\,d\theta,$$

whence

$$
\begin{aligned}
v(1,\phi) &= \tfrac{1}{2\pi}\int_0^{\alpha(\varepsilon)} \ln\left(\frac{\sqrt{H(1,\phi,\theta)}+\sqrt{M(1,\phi)-N(1,\phi)}+\sqrt{K(1,\phi,\theta)}}{\sqrt{H(1,\phi,\theta)}-\sqrt{M(1,\phi)-N(1,\phi)}+\sqrt{K(1,\phi,\theta)}}\right)\left|f'(e^{i\theta})\right|\,d\theta \\
&= \tfrac{1}{\pi}\int_0^{\alpha(\varepsilon)} \ln\left|\frac{\sqrt{\sin((\alpha(\varepsilon)-\theta)/2)\sin(\phi/2)}+\sqrt{\sin(\theta/2)\sin((\alpha(\varepsilon)-\phi)/2)}}{\sqrt{\sin((\alpha(\varepsilon)-\theta)/2)\sin(\phi/2)}-\sqrt{\sin(\theta/2)\sin((\alpha(\varepsilon)-\phi)/2)}}\right|\left|f'(e^{i\theta})\right|\,d\theta
\end{aligned}
\tag{49}
$$

and

$$
\begin{aligned}
\frac{\partial v}{\partial \rho}(1,\phi) &= -\frac{\sin(\alpha(\varepsilon)/2)}{2\pi\sqrt{\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}} \\
&\quad\times\int_0^{\alpha(\varepsilon)} \frac{\left|f'(e^{i\theta})\right|d\theta}{\sin\frac{\phi}{2}\sqrt{\frac{\sin((\alpha(\varepsilon)-\theta)/2)}{\sin(\theta/2)}}+\sin\frac{\phi-\alpha(\varepsilon)}{2}\sqrt{\frac{\sin(\theta/2)}{\sin((\alpha(\varepsilon)-\theta)/2)}}}.
\end{aligned}
\tag{50}
$$

Compare (49) and (50) with (35) and (41), respectively. Now we prove that a result like Theorem 7 holds for Jordan domains with a Dini-smooth boundary.

THEOREM 8. *Let Ω be a Jordan domain with a Dini-smooth boundary and Γ_1 denote a connected arc of $\partial\Omega$ such that $|\Gamma_1| = \varepsilon$. If $u[\Gamma_1]$ is the solution to problem (1) for Ω and Γ_1, then we have*

$$
\sup_\Omega u[\Gamma_1] = \tfrac{1}{2}\varepsilon + o(\varepsilon). \tag{51}
$$

Moreover, if the conformal map $f : B_1(0) \to \Omega$ and $\alpha(\varepsilon)$ are defined as above, then for $\alpha(\varepsilon) < \phi_1 < \phi_2 < 2\pi$,

$$
\sup_{\phi_1<\phi<\phi_2}\left|\frac{\partial u[\Gamma_1]}{\partial n}(f(e^{i\phi}))\right| \le \frac{\|f'\|_\infty}{8\pi\, M^3(f')\min_{\phi\in[\phi_1,\phi_2]}[\sin(\phi/2)\sin((\phi-\alpha)/2)]}\varepsilon^2 + o(\varepsilon^2), \tag{52}
$$

where $M(f') = \min_{0\le\phi\le 2\pi}|f'(\phi)|$.

Proof. By making an appropriate change of variable in the integral of the last member of (49), for $0 < \phi < \alpha(\varepsilon)$ we obtain

$$
\begin{aligned}
v(1,\phi) &= \tfrac{\alpha(\varepsilon)}{\pi}\int_0^1 \ln\left|\frac{\sqrt{\sin(\alpha(\varepsilon)(1-\lambda)/2)\sin(\phi/2)}+\sqrt{\sin(\alpha(\varepsilon)\,\lambda/2)\sin((\alpha(\varepsilon)-\phi)/2)}}{\sqrt{\sin(\alpha(\varepsilon)(1-\lambda)/2)\sin(\phi/2)}-\sqrt{\sin(\alpha(\varepsilon)\,\lambda/2)\sin((\alpha(\varepsilon)-\phi)/2)}}\right| \\
&\quad\times\left|f'(e^{i\alpha(\varepsilon)\,\lambda})\right|d\lambda.
\end{aligned}
\tag{53}
$$

Taking into account (48) and the continuity up to the boundary of f', we realize that, when $\varepsilon \downarrow 0$, the integral of the right-hand side of (53) converges to

$$
\left|f'(1)\right|\int_0^1 \ln\left|\frac{\sqrt{1-\lambda}+\sqrt{\lambda}}{\sqrt{1-\lambda}-\sqrt{\lambda}}\right|d\lambda = \tfrac{\pi}{2}\,\left|f'(1)\right|;
$$

thus

$$
\begin{aligned}
u[\Gamma_1](f(e^{i\phi})) &= v(1,\phi) \\
&= \tfrac{1}{\pi}\left(\tfrac{1}{|f'(1)|}\varepsilon + o(\varepsilon)\right)\left(\tfrac{\pi}{2}\left|f'(1)\right| + O(1)\right) = \tfrac{1}{2}\varepsilon + o(\varepsilon),
\end{aligned}
$$

so proving the expression (51). On the other hand, from (50) we deduce

$$\begin{aligned}\frac{\partial u[\Gamma_1]}{\partial n}(f(e^{i\phi})) &= \frac{1}{|f'(e^{i\phi})|}\frac{\partial v}{\partial \rho}(1,\phi)\\ &= -\frac{\sin(\alpha(\varepsilon)/2)}{2\pi|f'(e^{i\phi})|\sqrt{\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}}\\ &\times\alpha(\varepsilon)\int_0^1 \frac{\left|f'(e^{i\alpha(\varepsilon)\,\lambda})\right|d\lambda}{\sin\frac{\phi}{2}\sqrt{\frac{\sin(\alpha(\varepsilon)\,(1-\lambda)/2)}{\sin(\alpha(\varepsilon)\,\lambda/2)}}+\sin\frac{\phi-\alpha(\varepsilon)}{2}\sqrt{\frac{\sin(\alpha(\varepsilon)\,\lambda/2)}{\sin(\alpha(\varepsilon)\,(1-\lambda)/2)}}}.\end{aligned} \tag{54}$$

By proceeding as in the proof of (43) we obtain

$$\begin{aligned}\left|\frac{\partial u[\Gamma_1]}{\partial n}(f(e^{i\phi}))\right| &\le \frac{\alpha(\varepsilon)\sin(\alpha(\varepsilon)/2)}{2\pi\,|f'(e^{i\phi})|\sqrt{\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}}\,\frac{\int_0^1\left|f'(e^{i\alpha(\varepsilon)\,\lambda})\right|d\lambda}{2\sqrt{\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}}\\ &\le \frac{\left\|f'\right\|_\infty\alpha(\varepsilon)\sin(\alpha(\varepsilon)/2)}{4\pi\,|f'(e^{i\phi})|\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}\\ &= \frac{\left\|f'\right\|_\infty}{8\pi\,|f'(e^{i\phi})|\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}\,\frac{\varepsilon^2}{\left|f'(1)\right|^2}+o(\varepsilon^2)\\ &\le \frac{\left\|f'\right\|_\infty}{8\pi\,M^3(f')\sin(\phi/2)\sin((\phi-\alpha(\varepsilon))/2)}\varepsilon^2+o(\varepsilon^2),\end{aligned} \tag{55}$$

where we have again used (48), the continuity up to the boundary of f', and the fact that f' does not vanish in $\overline{\Omega}$. Inequality (52) follows by taking $\sup_{\phi_1<\phi<\phi_2}$ in the first and last members of inequalities (55). □

As a simple example of an application of Theorem 8 we now set $\Omega = B_R(0)$. The function $f(z) = iR\,(1-z)$ is the Riemann mapping from $B_1(0)$ onto $iR + B_R(0)$ verifying $f(1) = 0,\ f'(1) = -iR = -i\,|f'(1)|$. Since $|f'(z)| \equiv R$ for this mapping, expressions (45) and (46) are respectively recovered from the general expressions (51) and (52) appearing in the theorem.

Observe that expression (51) from Theorem 8 is just a proof of (26) for Jordan domains with a Dini-smooth boundary. In order to derive the corresponding expression (27) for these domains, we introduce suitable coordinates so that $a_1 = 0$, $\gamma(0) = O$, and $\gamma'(0) = 1$. As before, denote by f the function that conformally maps $B_1(0)$ onto Ω and satisfies $f(1) = 0,\ f'(1) = -i\,|f'(1)|$. For $k = 1, 2, \ldots, n$ and $0 < \varepsilon \le 1$, define the rotations $g_k(z) = e^{i\alpha(a_k(\varepsilon))}z$ and consider the mappings $f_k : B_1(0) \to \Omega$ such that $f_k = f \circ g_k$. These mappings verify $f_k(1) = f(e^{i\alpha(a_k(\varepsilon))}) = \gamma(a_k(\varepsilon))$, $f_k\left(e^{i(\alpha(b_k(\varepsilon))-\alpha(a_k(\varepsilon)))}\right) = f(e^{i\alpha(b_k(\varepsilon))}) = \gamma(b_k(\varepsilon))$, $\|f_k'\|_\infty = \|f'\|_\infty$, and $M(f_k') = M(f')$; then, an application of the inequality (52) with $u[\Gamma_1]$ and f substituted by v_k and f_k, respectively, give us

$$\sup_{\phi_1<\phi<\phi_2}\left|\frac{\partial v_k}{\partial n}(f_k(e^{i\phi}))\right| \le \frac{\left\|f'\right\|_\infty((b_k-a_k)\varepsilon)^2}{8\pi\,M^3(f')\min_{\phi\in[\phi_1,\phi_2]}\left[\sin\frac{\phi}{2}\sin\frac{\phi-\left(\alpha(b_k(\varepsilon))-\alpha(a_k(\varepsilon))\right)}{2}\right]}+o(\varepsilon^2),$$

where $\alpha(b_k(\varepsilon)) - \alpha(a_k(\varepsilon)) < \phi_1 < \phi_2 < 2\pi$. Therefore, for small enough ε 's, we have

$$\left\|\left(\partial v_k/\partial n\right)|_{\Gamma_1^{(j)}(\varepsilon)}\right\|_\infty \le \frac{\left\|f'\right\|_\infty((b_k-a_k)\varepsilon)^2}{8\pi\,M^3(f')\min_{\phi\in[\alpha(a_j(\varepsilon)),\alpha(b_j(\varepsilon))]}\left[\sin\frac{\phi-\alpha(a_k(\varepsilon))}{2}\sin\frac{\phi-\alpha(b_k(\varepsilon))}{2}\right]}+o(\varepsilon^2). \tag{56}$$

Since

$$\min_{\phi\in[\alpha(a_j(\varepsilon)),\alpha(b_j(\varepsilon))]}\left[\sin\tfrac{\phi-\alpha(a_k(\varepsilon))}{2}\sin\tfrac{\phi-\alpha(b_k(\varepsilon))}{2}\right]\to\sin^2\left(\tfrac{\alpha((a_j+b_j)/2)-\alpha((a_k+b_k)/2)}{2}\right)$$

when $\varepsilon\downarrow 0$, we see from (56) that expression (27) holds for the constants $C_{j,k}$ given by

$$C_{j,k}=\frac{\|f'\|_\infty}{4\pi\, M^3(f')\min_{j\neq k}\left[\sin^2\left(\frac{\alpha((a_j+b_j)/2)-\alpha((a_k+b_k)/2)}{2}\right)\right]}.$$

4. Proof of Theorem 5. In order to prove Theorem 5, better estimates for normal derivatives are needed. For the sake of clarity, we first discuss the case $\Omega=B_1(0)$. Notations are the same as those used in the previous section.

From (41) and (44) we obtain, for $\alpha<\phi<2\pi$,

$$\begin{aligned}\left|\tfrac{\partial u_\alpha}{\partial\rho}(1,\phi)\right|&\le\tfrac{\alpha\sin(\alpha/2)}{4\pi\sin(\phi/2)\sin((\phi-\alpha)/2)}\\&=\tfrac{\alpha}{4\pi}\tfrac{\sin(\phi/2)\cos((\phi-\alpha)/2)-\cos(\phi/2)\sin((\phi-\alpha)/2)}{\sin(\phi/2)\sin((\phi-\alpha)/2)}\\&=\tfrac{\alpha}{4\pi}\left(\cot((\phi-\alpha)/2)-\cot(\phi/2)\right).\end{aligned}$$

By the mean value theorem, we can write

$$\cot((\phi-\alpha)/2)-\cot(\phi/2)=\sin^{-2}\left((\phi-\alpha)/2+\mu\alpha/2\right)\alpha/2$$

for a certain $0<\mu=\mu(\phi)<1$, and then

$$\left|\tfrac{\partial u_\alpha}{\partial\rho}(1,\phi)\right|\le\tfrac{\alpha^2}{8\pi\sin^2((\phi-\alpha)/2+\mu(\phi)\alpha/2)},\ \alpha<\phi<2\pi. \tag{57}$$

In this way, for a given $\tau>0$ we have

$$\begin{aligned}\sup_{\alpha+\tau<\phi<2\pi-\tau}\left|\tfrac{\partial u_\alpha}{\partial\rho}(1,\phi)\right|&\le\tfrac{\alpha^2}{8\pi\min_{\alpha+\tau<\phi<2\pi-\tau}[\sin^2((\phi-\alpha)/2+\mu(\phi)\alpha/2)]}\\&\le\tfrac{\alpha^2}{8\pi\sin^2(\tau/2)}.\end{aligned} \tag{58}$$

Now consider a finite family of arcs $\Gamma_1=\cup_{k=1}^n\Gamma_1^{(k)}$ with $\Gamma_1^{(k)}=(\alpha_k,\beta_k)\subset\partial B_1(0)$ and let δ denote the minimum distance on $\partial B_1(0)$ between adjacent components of Γ_1; i.e., $\delta=\min\{\min_{1\le k\le n-1}(\alpha_{k+1}-\beta_k),\alpha_1-\beta_n\}$. By using (58), we find

$$\left\|\left.\frac{\partial v_k}{\partial n}\right|_{\Gamma_1^{(j)}}\right\|_\infty\le\frac{(\beta_k-\alpha_k)^2}{8\pi\sin^2(\delta/2)}, \tag{59}$$

so that

$$\sum_{k=1}^n\left\|\left.\frac{\partial v_k}{\partial n}\right|_{\Gamma_1^{(j)}}\right\|_\infty\le\frac{1}{8\pi\sin^2(\delta/2)}\sum_{k=1}^n(\beta_k-\alpha_k)^2\le\frac{|\Gamma_1|^2}{8\pi\sin^2(\delta/2)}. \tag{60}$$

Now, inequalities like (59) and (60) will be analogously derived for a domain with a Dini-smooth boundary. With this purpose, the starting point we will choose is a general representation formula for the normal derivative of the solution to the problem

$$\left|\begin{array}{ll}\Delta u(\rho,\phi)=\frac{1}{\rho}\left((\rho u_\rho)_\rho+\frac{1}{\rho}\,u_{\phi\phi}\right)=0, & (\rho,\phi)\in B_1(0),\\ u(1,\phi)=0, & \beta<\phi<\alpha+2\pi,\\ \frac{\partial u}{\partial\rho}(1,\phi)=\left|f'(e^{i\phi})\right|, & \alpha<\phi<\beta,\end{array}\right.$$

where the function f conformally maps $B_1(0)$ onto Ω. Denote by $u_{\alpha,\beta}$ the solution to this problem. After the developments of the previous section, it is not difficult to see that, for $\beta < \phi < \alpha + 2\pi$,

$$\frac{\partial u_{\alpha,\beta}}{\partial \rho}(1,\phi) = -\frac{\sin((\beta-\alpha)/2)}{2\pi\sqrt{\sin\frac{\phi-\alpha}{2}\sin\frac{\phi-\beta}{2}}}\int_\alpha^\beta \frac{\left|f'(e^{i\theta})\right|d\theta}{\sin\frac{\phi-\alpha}{2}\sqrt{\frac{\sin((\beta-\theta)/2)}{\sin((\theta-\alpha)/2)}}+\sin\frac{\phi-\beta}{2}\sqrt{\frac{\sin((\theta-\alpha)/2)}{\sin((\beta-\theta)/2)}}},$$

whence we deduce

$$\begin{aligned}\frac{\partial v_k}{\partial n}(f(e^{i\phi})) &= -\frac{\sin((\beta_k-\alpha_k)/2)}{2\pi\left|f'(e^{i\phi})\right|\sqrt{\sin\frac{\phi-\alpha_k}{2}\sin\frac{\phi-\beta_k}{2}}}\\ &\quad\times\int_{\alpha_k}^{\beta_k}\frac{\left|f'(e^{i\theta})\right|d\theta}{\sin\frac{\phi-\alpha_k}{2}\sqrt{\frac{\sin((\beta_k-\theta)/2)}{\sin((\theta-\alpha_k)/2)}}+\sin\frac{\phi-\beta_k}{2}\sqrt{\frac{\sin((\theta-\alpha_k)/2)}{\sin((\beta_k-\theta)/2)}}}.\end{aligned} \tag{61}$$

In estimating the integral in the second member of (61) we again use the arithmetic mean-geometric mean inequality and the fact that $\int_\alpha^\beta \left|f'(e^{i\theta})\right| d\theta = b - a$ to obtain

$$\begin{aligned}\left|\frac{\partial v_k}{\partial n}(f(e^{i\phi}))\right| &\le \frac{\sin((\beta_k-\alpha_k)/2)}{4\pi\left|f'(e^{i\phi})\right|\sin\frac{\phi-\alpha_k}{2}\sin\frac{\phi-\beta_k}{2}}\int_{\alpha_k}^{\beta_k}\left|f'(e^{i\theta})\right|d\theta\\ &= \frac{b_k-a_k}{4\pi\left|f'(e^{i\phi})\right|}\frac{\sin\frac{\beta_k-\alpha_k}{2}}{\sin\frac{\phi-\alpha_k}{2}\sin\frac{\phi-\beta_k}{2}}.\end{aligned} \tag{62}$$

As before, we have

$$\frac{\sin\frac{\beta_k-\alpha_k}{2}}{\sin\frac{\phi-\alpha_k}{2}\sin\frac{\phi-\beta_k}{2}} = \frac{\beta_k-\alpha_k}{2\sin^2[\phi-((1-\mu)\alpha_k+\mu\beta_k))]},$$

for $0 < \mu = \mu(\phi) < 1$, and then inequality (62) becomes

$$\left|\frac{\partial v_k}{\partial n}(f(e^{i\phi}))\right| \le \frac{b_k-a_k}{8\pi\left|f'(e^{i\phi})\right|}\frac{\beta_k-\alpha_k}{\sin^2[\phi-((1-\mu)\alpha_k+\mu\beta_k))]}. \tag{63}$$

Now let d be the minimum distance between adjacent components of Γ_1 as defined by (8). Since $M(f') \le \left|f'(e^{i\theta})\right| \le \|f'\|_\infty$, we have

$$M(f')\delta \le d \le \|f'\|_\infty \delta,$$

where $\delta = \min\{\min_{1\le k\le n-1}(\alpha_{k+1}-\beta_k), \alpha_1-\beta_n\}$ is the minimum distance on $\partial B_1(0)$ between adjacent components of $f^{-1}(\Gamma_1)$. Thus, from (62) we deduce

$$\begin{aligned}\left\|\frac{\partial v_k}{\partial n}\Big|_{\Gamma_1^{(j)}}\right\|_\infty &\le \sup_{\alpha_k-\delta<\phi<\beta_k+\delta}\left|\frac{\partial v_k}{\partial n}(f(e^{i\phi}))\right|\\ &\le \sup_{\alpha_k-\delta<\phi<\beta_k+\delta}\left[\frac{b_k-a_k}{8\pi\left|f'(e^{i\phi})\right|}\frac{\beta_k-\alpha_k}{\sin^2[\phi-((1-\mu)\alpha_k+\mu\beta_k)]}\right]\\ &\le \frac{\|f'\|_\infty(b_k-a_k)^2}{8\pi M(f')}\sup_{\alpha_k-d/\|f'\|_\infty<\phi<\beta_k+d/\|f'\|_\infty}\left[\frac{1}{\sin^2[\phi-((1-\mu)\alpha_k+\mu\beta_k)]}\right]\\ &\le \frac{\|f'\|_\infty(b_k-a_k)^2}{8\pi M(f')}\frac{1}{\sin^2\left[d/\left(2\|f'\|_\infty\right)\right]},\end{aligned} \tag{64}$$

and summing on $k = 1, 2, \ldots, n$,

$$\sum_{k=1}^{n} \left\| \frac{\partial v_k}{\partial n}\Big|_{\Gamma_1^{(j)}} \right\|_\infty \leq \frac{\|f'\|_\infty}{8\pi\, M(f')} \frac{1}{\sin^2\left[d/\left(2\|f'\|_\infty\right)\right]} \sum_{k=1}^{n} (b_k - a_k)^2$$

$$\leq \frac{\|f'\|_\infty}{8\pi\, M(f')} \frac{|\Gamma_1|^2}{\sin^2\left[d/\left(2\|f'\|_\infty\right)\right]} \tag{65}$$

$$= \frac{\|f'\|_\infty}{8\pi\, M(f')} H^2(\Gamma_1),$$

where $H(\Gamma_1)$ was defined by (9).

Finally, with inequalities (64) and (65) at hand, we can proceed to prove Theorem 5 as follows. Assume that $\Gamma_1 \in \mathcal{F}$; i.e., that there exist δ_1 and δ_2, $0 < \delta_1 < \delta_2 < 1$, such that

$$\rho(\Gamma_1) \leq \delta_1, \tag{66}$$

$$\frac{\|f'\|_\infty}{8\pi\, M(f')} H^2(\Gamma_1) \leq 1 - \delta_2. \tag{67}$$

Furthermore, take $0 < \varepsilon \leq (\delta_2 - \delta_1) / [2(\delta_2 + \delta_1)]$ and assume that $|\Gamma_1|$ is so small that expression (51) from Theorem 8 can be applied to write

$$\max_{1\leq k\leq n} \sup_\Omega v_k < \frac{1}{2} \max_{1\leq k\leq n} \left|\Gamma_1^{(k)}\right| + \varepsilon \max_{1\leq k\leq n} \left|\Gamma_1^{(k)}\right| \tag{68}$$

and

$$\sup_\Omega u[\Gamma_1^*] > \frac{1}{2} |\Gamma_1| - \varepsilon |\Gamma_1|, \tag{69}$$

where $\Gamma_1^* \subset \partial\Omega$ is an arc with $|\Gamma_1^*| = |\Gamma_1|$. From inequalities (65) and (67) we see that the second inequality in Lemma 6(iv) is applicable so that, in view of (68), (66), and (67), we obtain

$$\begin{aligned}
\sup_\Omega u[\Gamma_1] &\leq \frac{\max_{1\leq k\leq n} \sup_\Omega v_k}{1 - \sum_{k=1}^{n} \left\| (\partial v_k/\partial n)|_{\Gamma_1^{(j)}} \right\|_\infty} \\
&< \frac{\frac{1}{2} \max_{1\leq k\leq n} \left|\Gamma_1^{(k)}\right| + \varepsilon \max_{1\leq k\leq n} \left|\Gamma_1^{(k)}\right|}{1 - \frac{\|f'\|_\infty}{8\pi\, M(f')} H^2(\Gamma_1)} \\
&= \frac{\frac{1}{2} + \varepsilon}{1 - \frac{\|f'\|_\infty}{8\pi\, M(f')} H^2(\Gamma_1)} \rho(\Gamma_1) |\Gamma_1| \\
&\leq \frac{\delta_1}{\delta_2} \left(\frac{1}{2} + \varepsilon\right) |\Gamma_1|. \qquad (70)
\end{aligned}$$

From (70) and (69) we deduce

$$\sup_\Omega u[\Gamma_1] < \frac{\delta_1}{\delta_2} \left(\frac{1}{2} + \varepsilon\right) |\Gamma_1| \leq \left(\frac{1}{2} - \varepsilon\right) |\Gamma_1| < \sup_\Omega u[\Gamma_1^*].$$

This finishes the proof of Theorem 5.

5. Final remarks. Results that are to Conjectures 2 and 3 as Theorem 4 is to Conjecture 1 can reasonably be proved. Namely, if Ω is a sufficiently regular Jordan domain, $P \in \Omega$ and $1 \leq p < +\infty$, then the inequalities

$$\|u[\Gamma_1(\varepsilon)]\|_p < \|u[\Gamma_1^*(\varepsilon)]\|_p ,$$
$$u[\Gamma_1(\varepsilon)](P) < u[\Gamma_1^*(\varepsilon)](P),$$

hold for ε small enough, provided that Γ_1 and Γ_1^* are two subsets of γ respectively given by (3) and (5) with $\beta - \alpha = \sum_{k=1}^{n}(\beta_k - \alpha_k)$.

Although questions of regularity of the domain Ω may well be considered of secondary interest in relation to the proposed conjectures, some commentaries are in order on this matter. Domains Ω with a Dini-smooth boundary were assumed in the developments of sections 4 and 5. By eventually restricting the placement on $\partial\Omega$ of the arcs composing Γ_1, this assumption may be relaxed to include domains with corners too. The technique of conformal maps could be employed again in such a generalization (cf. [12]). On the other side, the domain Ω studied in section 2 is one with corners, but these corners can be mollified to obtain a smooth domain $\Omega_* \supset \Omega$ so close to Ω that the solution u_* corresponding to problem (4) for Ω_* (and an appropriate Γ_1) is a small "perturbation" of the solution u to that problem (4) for Ω. This domain Ω_* would then afford an example of a nonconvex smooth domain for which the optimality of connected arcs is not true.

In this paper we have treated plane domains but n-dimensional versions of Conjectures 1–3 could, we hope, also be supportable. Of course, additional precisions will then be needed on the geometry of the admissible family Γ_1. For the sphere, the optimal Γ_1 would presumably be bounded by circles.

Acknowledgments. The author is gratefully indebted to Prof. Luis A. Caffarelli for his many useful observations and valuable suggestions on the subject of this paper. He also wishes to thank the hospitality of the Courant Institute of Mathematical Sciences, New York University, where the research conducted for this paper was carried out. Finally, the author expresses his gratitude to the anonymous referee whose clever comments and criticism helped this paper to attain the present form.

REFERENCES

[1] S. Axler, P. Bourdon, and W. Ramey, *Harmonic Function Theory*, Springer, New York, 1991.

[2] L. R. Berrone, *Subsistencia de Modelos Matematicos que Involucran a la Ecuacion del Calor-Difusion*, Ph.D. thesis, Universidad Nacional de Rosario, Argentina, 1994.

[3] L. R. Berrone, *Explicit bounds for harmonic functions satisfying boundary conditions of mixed type*, Canad. Appl. Math. Quart., 5 (1997), pp. 171–204.

[4] L. R. Berrone, *On a conjecture relative to the maximum of harmonic functions on convex domains: Unbounded domains*, Portugal. Math., 55 (1998), pp. 307–321.

[5] L. R. Berrone, *Lifting a circular membrane by unitary forces*, Glas. Mat. Ser. III, 34 (1999), pp. 5–10.

[6] U. Dini, *Sur la méthode des approximations successives pour les équations aux derivées partielles du deuxième ordre*, Acta Math., 25 (1902), pp. 185–230.

[7] A. Ghizzetti, *Sopra un particolare problema misto di Dirichlet-Neumann per l'equazione di Laplace, trattato col metodo delle trasformati parziali*, Rend. Mat. Appl., 5 (1946), pp. 131–168.

[8] A. Ghizzetti, *Sopra due particolari problemi misti di Dirichlet-Neumann per l'equazione di Laplace*, Rend. Accad. Naz. Lincei, Serie VIII, vol. I, 1 (1946), pp. 40–44.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer, Berlin, 1983.

[10] O. D. Kellogg, *Potential functions on the boundary of their regions of definitions*, Trans. Amer. Math. Soc., 9 (1908), pp. 39–50.
[11] E. Liban, Ph.D. *thesis*, New York University, NY, 1957.
[12] Ch. Pommerenke, *Boundary Behaviour of Conformal Maps*, Springer, Berlin, 1991.
[13] M. H. Protter and H. F. Weinberger, *Maximum Principles in Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1967.
[14] I. N. Sneddon, *Mixed Boundary Problems in Potential Theory*, North–Holland, Amsterdam, 1966.
[15] W. L. Wendland, E. Stephan, and G. C. Hsiao, *On the integral equation method for the plane mixed boundary problem of the laplacian*, Math. Methods Appl. Sci., 1 (1979), pp. 265–321.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1208–1240

FIRST-ORDER CORRECTOR FOR THE HOMOGENIZATION OF THE CRITICALITY EIGENVALUE PROBLEM IN THE EVEN PARITY FORMULATION OF THE NEUTRON TRANSPORT*

GUILLAUME BAL†

Abstract. We consider the homogenization of the criticality eigenvalue problem for the even parity flux of neutron transport in a domain with isotropic and periodically oscillating coefficients. We prove that the neutron density is factored in the product of two terms. The first one describes local behavior of the density at the cell level. It is a solution of a heterogeneous transport problem with periodic boundary conditions. The second term gives global behavior on the whole domain. It satisfies a homogeneous diffusion equation posed on the whole domain with Dirichlet boundary conditions. We also give the asymptotic analysis of the corresponding eigenvalues. This expansion gives rise to errors of the order of the cell size. It does not account for neutron leakage at the boundary of the core and yields unacceptable errors in practice. We derive a more accurate expansion of the eigenelements in the case of a symmetric and cubic domain. The analysis of a boundary layer allows us to derive modified boundary conditions for the diffusion eigenvalue problem. The resulting approximation for the leading transport eigenvalue is proven to be accurate to one order higher than previously. Numerical experiments confirm the accuracy of the reconstructed eigenvectors in realistic settings.

Key words. neutron transport, eigenvalue problem, homogenization, even parity flux formulation, half space problem

AMS subject classification. 35B27

PII. S0036141098338855

1. Introduction.

Transport equations are solved in industrial applications in order to determine the power distribution of neutrons in nuclear reactors. In the case of a stable reactor, only a steady-state solution is required. Hence the time variable can be eliminated. An eigenvalue problem, called the criticality problem for neutron transport, is solved to figure out whether a steady state exists. The unknowns are the multiplicative factor k_{eff}, which expresses the balance between the production of neutrons by fission and its absorption and leakage at the boundary of the core, and the neutron density $\phi(x,\mu)$, defined in phase space at position x and velocity μ. They are the largest eigenvalue and the associated positive eigenvector of the following equation:

$$(1)\quad (\mu\cdot\nabla\phi+\Sigma\phi)(x,\mu)=\int_V \Sigma_s(x,\mu',\mu)\phi(x,\mu')d\mu'+\frac{1}{k_{eff}}\int_V \sigma_f(x,\mu',\mu)\phi(x,\mu')d\mu'$$

posed with appropriate boundary conditions in an open bounded domain $\Omega\in\mathbb{R}^d$ for a velocity space V. With the usual notation, $\mu\cdot\nabla=\sum_{i=1}^d \mu_i\frac{\partial}{\partial x_i}$, where $\mu=(\mu_1,\dots,\mu_d)$ and $d\in\mathbb{N}^*$ is the spatial dimension. It turns out that the eigenvector $\phi(x,v)$ is the only positive normalized eigenvector of (1); hence the only one of physical interest. We have the following interpretation for the multiplicative factor k_{eff}. If $k_{eff}=1$, the core is stable and fission exactly compensates for absorption and leakage. When

*Received by the editors May 18, 1998; accepted for publication (in revised form) January 5, 1999; published electronically October 4, 1999.
http://www.siam.org/journals/sima/30-6/33885.html

†Électricité de France DER/IMA/MMN, 92141 Clamart, France, and Laboratoire d'Analyse Numérique, Université Paris-VI, 75252 Paris Cedex 5, France. Current address: Department of Mathematics, Stanford University, Stanford, CA 94305 (bal@math.stanford.edu).

$k_{eff} > 1$, fission is too important and the reactor is supercritical. If $k_{eff} < 1$, fission must be increased, or the chain reaction dies out.

In (1), Σ, Σ_s, and σ_f are the total cross section, the scattering cross section, and the fission cross section, respectively. They characterize the nuclear reactors, which are highly heterogeneous. Therefore, numerical simulations with the transport equation are very demanding.

A first approximation consists in assuming that the core is periodic. It allows us to homogenize the transport equation (1), which yields a homogeneous second-order elliptic eigenvalue problem. We consider here a physical domain Ω^ε composed of roughly $\frac{1}{\varepsilon^d}$ identical cells. The cross sections Σ, Σ_s, and σ_f are supposed to be isotropic, i.e., only depend on the spatial position x, and to be Y-periodic in the domain Ω^ε, where Y is the unit cell. For simplicity, the velocity space V will be here the unit sphere $V = S^{d-1} = \{\mu \in \mathbb{R}^d,\ |\mu| = 1\}$.

The homogenization of transport equations has been studied at length in the past, physically [12, 15, 21] and mathematically [14, 26, 27, 31, 33]. We have recently revisited the criticality eigenvalue problem in [4, 5, 8]. To the best of our knowledge, no theory is available for the homogenization of eigenvalue problems with periodic coefficients taking account for neutron leakage at the boundary of the core. In the same spirit, let us mention the works of [6, 7, 29] concerning the homogenization of heterogeneous diffusion eigenvalue problems. Notice that various results on the homogenization of transport equations have been obtained in different contexts [1, 22, 23].

The theory of transport is usually done using (1). However, it is also interesting to analyze the so-called even parity flux formulation. The *even parity flux* is defined by

$$\psi^+(x,\mu) = \frac{1}{2}(\phi(x,\mu) + \phi(x,-\mu)).$$

When no direction is privileged with respect to its opposite one, such as in the diffusion limit, this symmetrized flux can be of interest. We deduce from (1) and Appendix A the following equation in $\Omega^\varepsilon \times V$:

$$-\mu\nabla\frac{1}{\Sigma(x)}\mu\nabla\psi^+(x,\mu)+\Sigma(x)\psi^+(x,\mu) = \left(\Sigma_s(x) + \frac{1}{k^\varepsilon_{eff}}\sigma_f(x)\right)\int_V \psi^+(x,\mu')d\mu. \tag{2}$$

The derivation of this symmetrized transport equation requires that the cross sections be isotropic. By definition of the even parity flux, we have $\psi^+(x,-\mu) = \psi^+(x,\mu)$. Consequently, we need to consider only solutions of (2) satisfying this symmetry condition. It is well known that the homogenization of transport equations yields homogeneous second-order elliptic equations. Thus, the second-order differential operator in the even parity formulation is similar to that of diffusion. More precisely, we will see that the variational formulation associated with the transport equation (2) is well suited to the derivation of the variational formulation of diffusion and simplifies the analysis of the leakage at the boundary of the core and of the associated boundary layer problem.

Let us now define our main framework. It is convenient to recast the sequence of problems (2), parameterized by ε, on a fixed domain. By change of variables $x \to \varepsilon x$, the spatial domain becomes Ω. Since the core is periodic, Ω is composed of roughly ε^{-d} identical cells of the kind $\varepsilon Y = (0,\varepsilon)^d$, where $Y = (0,1)^d$ is the unit cell. The

period of the heterogeneities in Ω is thus given by $\varepsilon > 0$. No further geometrical assumption is needed when no leakage is accounted for. However, in order to derive a more accurate approximation for the leading eigenvalue, neutron leakage at the boundary cannot be ignored. We have to make more precise the local geometry at the boundary of the core. We assume here that Ω is the unit cube and that ε^{-1} is an integer. The main characteristic is that the boundary of Ω always coincides with the boundary of a unit cell.

Let us define

$$\Sigma^\varepsilon(x) = \Sigma\left(\frac{x}{\varepsilon}\right), \Sigma_s^\varepsilon(x) = \Sigma_s\left(\frac{x}{\varepsilon}\right), \sigma_f^\varepsilon(x) = \sigma_f\left(\frac{x}{\varepsilon}\right), \tag{3}$$

where Σ, Σ_s, and σ_f are Y-periodic measurable functions, bounded from above and below by positive constants and such that the absorption cross section $\Sigma - \Sigma_s$ be also bounded from below by a positive constant. The latter condition ensures that absorption is positive everywhere; hence fission is allowed. For simplicity, we introduce $\lambda_\varepsilon = \frac{1}{k_{eff}}$. We consider here an absorbing boundary condition. It means that no particle enters the core. The solution to (1) satisfies

$$\phi = 0 \text{ on } \Gamma_- = \{(x,\mu) \in \partial\Omega \times V \text{ subject to (s.t.) } \mu \cdot n(x) < 0\},$$

where $n(x)$ denotes the outward unit normal to $\partial\Omega$ at $x \in \partial\Omega$. In practice, this boundary condition is a first approximation only. The addition of a reflector around the core yields a correction of order ε. Its analysis is much more involved and is not considered here.

Let us define $\varphi^{\varepsilon+}(x,\mu) = \psi^+(\frac{x}{\varepsilon},\mu)$. These assumptions enable us to recast (2) as

$$\begin{aligned} &-\varepsilon^2 \mu\cdot\nabla \frac{1}{\Sigma^\varepsilon(x)} \mu\cdot\nabla\varphi^{\varepsilon+}(x,\mu) + \Sigma^\varepsilon(x)\varphi^{\varepsilon+}(x,\mu) \\ &\qquad = (\Sigma_s^\varepsilon(x) + \lambda^\varepsilon \sigma_f^\varepsilon(x)) \int_V \varphi^{\varepsilon+}(x,\mu')d\mu' \quad \text{in } \Omega \times V, \\ &\varphi^{\varepsilon+}(x,\mu) - \frac{\varepsilon}{\Sigma^\varepsilon(x)} \mu\cdot\nabla\varphi^{\varepsilon+}(x,\mu) = 0 \quad \text{on } \Gamma_-. \end{aligned} \tag{4}$$

The boundary conditions have been obtained following rules recalled in Appendix A.

An outline of this paper is as follows. In section 2, we recall existing results on problem (4) at ε given. We also address the transport problem with periodic boundary conditions, which characterizes the neutron density at the small scale. In section 3, we present our main results on the asymptotic behavior of the eigenvalues and eigenvectors of (4) as $\varepsilon \to 0$. First an analysis is given without accounting for the leakage at the boundary of the core. Second, a first-order correction of the previous results is given for the largest eigenvalue k_{eff}. This correction characterizes the amount of neutron leakage at the boundary of a nuclear reactor. We give in sections 4 to 6 a detailed proof of these results. In section 4, we give some a priori estimates and existence results with ε fixed. Our results are based on the analysis of a source problem introduced in section 3. The asymptotic expansion of this problem is given in section 5. The analysis of a genuine multidimensional boundary layer problem, or Milne problem, that is used in section 5, is given in section 6. Finally, we present some numerical experiments in section 7.

2. The criticality eigenvalue problem in bounded and periodic domain. In this section, we state some results of existence and regularity for the eigenvalues and eigenvectors of the even parity transport. They are very close to known results

on eigenvalue problems in transport theory (see [20, Chapter 21]). We recall them here for the sake of completeness. The proofs are sketched only and we refer to the thesis [8] for the details and to the equivalence between the first-order formulation and the even parity flux formulation given in Appendix A. Our main hypothesis on the physical data is given now.

HYPOTHESIS 2.1.

(H1) Ω *is a convex open bounded subset of* $\mathbb{R}^d$.

(H2) $V = S^{d-1} = \{v \in \mathbb{R}^d \text{ s.t. } |v| = 1\}$.

(H3) *The cross sections* $\Sigma(x)$, $\Sigma_s(x)$, *and* $\sigma_f(x)$ *are positive functions in* $L^\infty(\Omega)$. *Moreover they are* Y*-periodic, where* $Y = (0,1)^d$ *is the periodicity cell.*

(H4) *There exists a constant* $\eta > 0$ *such that* $\Sigma(x) - \Sigma_s(x) \geq \eta$ *and* $\sigma_f(x) \geq \eta$.

Let us introduce the Hilbert space

$$W^2(\Omega \times V) = \{u \in L^2(\Omega \times V),\ \mu \cdot \nabla u \in L^2(\Omega \times V)\}. \tag{5}$$

We deduce from [8, Theorem II.2.1.1] and Appendix A the following result.

THEOREM 2.2. *Assume that Hypothesis* 2.1 *is satisfied. Then problem* (4) *admits a countable number of real eigenvalues and of associated eigenvectors, which are elements of* $W^2(\Omega \times V)$. *Moreover there exists a simple, positive, and real eigenvalue of smallest modulus, such that its associated eigenvector be the unique normalized positive eigenvector of* (4).

The solutions of (4) can be seen as the eigenvalues and eigenvectors of a positive compact operator. The reality of these eigenvalues, which holds true in the simplified setting of isotropic cross sections, is given in [34], for instance. The first part of the theorem is then proven. The second part relies on the Krein–Rutman theory of positive operators, which asserts that the spectral radius of this compact operator is an eigenvalue and that the corresponding eigenvector is positive. Following a proof given in [20, Chapter 21], one proves that this eigenvalue is simple and that there exists a unique positive normalized eigenvector.

As we shall see in the next section, the asymptotic limit as $\varepsilon \to 0$ of the solutions to (4) involves the small scale behavior of the neutron density. It is obtained by considering the solutions of the following criticality eigenvalue problem in periodic domain:

Find the smallest eigenvalue λ_∞ and the associated positive eigenvector ψ_∞^+ of

$$\begin{aligned} &-\mu \cdot \nabla_y \frac{1}{\Sigma} \mu \cdot \nabla_y \psi_\infty^+ + \Sigma \psi_\infty^+ \ = \ (\Sigma_s + \lambda_\infty \sigma_f) \int_V \psi_\infty^+(y, \mu') d\mu' \quad \text{in } Y \times V, \\ &y \mapsto \psi_\infty^+(y, \mu) \ \text{ is } Y\text{-periodic.} \end{aligned} \tag{6}$$

Still following the equivalence between the first-order formulation and the even parity flux formulation given in Appendix A, we deduce from [8, Theorems II.2.2.4 and II.2.2.7] the following theorem.

THEOREM 2.3. *Assume that Hypothesis* 2.1 *is satisfied. Then problem* (6) *admits a simple, positive, and real eigenvalue of smallest modulus, such that its associated eigenvector be the unique normalized positive eigenvector of* (6). *Moreover assume that the cross sections* Σ, Σ_s, *and* σ_f *belong to* $C^m(Y)$ *for* $m \in \mathbb{N}^*$. *Then we have that* $\psi_\infty^+ \in H^m(Y \times V)$.

REMARK 2.4. *The regularity result stated in the previous theorem is a characteristic property of the criticality problem with periodic boundary conditions. It is well known that the solutions to* (4) *are not arbitrarily regular, even with smooth physical data (see* [30]*). For example, we have a nonvanishing outgoing density at the boundary*

of the core, whereas the incoming density is zero by hypothesis. Let us also mention the following regularity result [5]. *Assume that the cross sections are bounded from above and below by positive constants. Then* ψ^+_∞ *and* $\frac{1}{\psi^+_\infty}$ *belong to* $L^\infty(Y\times V)$.

Consider the associated source problem:

$$\begin{aligned} &-\mu\cdot\nabla_y\frac{1}{\Sigma}\mu\cdot\nabla_y\psi^+ + \Sigma\psi^+ \;=\; (\Sigma_s+\lambda_\infty\sigma_f)\int_V \psi^+(y,\mu')d\mu' \;+\; S \quad \text{in } Y\times V,\\ &y\mapsto\psi^+(y,\mu) \;\text{ is } Y\text{-periodic,}\end{aligned} \tag{7}$$

where S is a given function in $L^2(Y\times V)$ satisfying $S(y,\mu)=S(y,-\mu)$. Then we deduce from [8, Theorems II.2.2.5 and II.2.2.8] the following result.

THEOREM 2.5. *Let* $(\lambda_\infty,\psi^+_\infty)$ *be the solution to* (6). *Then problem* (7) *admits a solution if and only if the source term* S *satisfies the following compatibility condition:*

$$\int_Y\int_V S(y,\mu)\psi^+_\infty(y,\mu)\,d\mu\,dy = 0. \tag{8}$$

Furthermore, if a solution exists, it is unique up to the addition of a multiple of ψ^+_∞. *Assume moreover that the cross sections* Σ, Σ_s, *and* σ_f *are of class* $C^m(Y)$ *and the source term* S *belongs to* $H^m(Y\times V)$ *for* $m\in\mathbb{N}^*$. *Then we have that* $\psi^+\in H^m(Y\times V)$.

3. Main results on the asymptotic analysis. In this section, we present our main results on the asymptotic analysis of the criticality eigenvalue problem (4). Let us introduce the function $\psi^+_\varepsilon(x,\mu)=\psi^+_\infty(\frac{x}{\varepsilon},\mu)$, where ψ^+_∞ is the positive normalized eigenvector of (6) extended by Y-periodicity on $\mathbb{R}^d\times V$. Clearly, ψ^+_ε is εY-periodic and satisfies (4) on $\Omega\times V$. The difference between ψ^+_ε and $\varphi^{\varepsilon+}$ is the definition of their boundary conditions on $\partial\Omega$. Therefore, we can expect some similarities in the behavior of both solutions away from the boundary. This is confirmed more precisely in Theorem 3.2 below. It is based on the asymptotic expansion of solutions to source problems, which requires sufficient regularity for the physical parameters. We do not dwell on optimal regularity results here and assume the physical parameters to be smooth.

HYPOTHESIS 3.1.

(H5) *In addition to Hypothesis* 2.1, *we assume the cross sections* Σ, Σ_s, *and* σ_f *to be of class* $C^\infty(\Omega)$.

(H6) *The domain* Ω *is either cubic or has a boundary* $\partial\Omega$ *of class* C^∞.

THEOREM 3.2. *Assume that Hypothesis* 3.1 *is satisfied. Let* $0<\lambda^\varepsilon_1<\lambda^\varepsilon_2\le\lambda^\varepsilon_3\le\cdots\le\infty$ *be the eigenvalues of* (4) *ranked in increasing order, and* $\varphi^{\varepsilon+}_l$ *the normalized eigenvector associated with* λ^ε_l. *Then, up to a subsequence, we have*

$$\varphi^{\varepsilon+}_l(x,\mu) \;=\; u_l(x)\psi^+_\infty(\frac{x}{\varepsilon},\mu)+O(\varepsilon) \quad \text{and} \quad \lambda^\varepsilon_l \;=\; \lambda_\infty+\varepsilon^2\nu_l+O(\varepsilon^3), \tag{9}$$

where λ_∞ *is the first eigenvalue of* (6), ψ^+_∞ *its associated normalized eigenvector and where* (ν_l,u_l) *are the* l*th eigenvalue (in the sense that* $0<\nu_1<\nu_2\le\nu_3\le\cdots\le\infty$*) and corresponding normalized eigenvector of the homogenized diffusion problem*

$$\begin{aligned} -\nabla D\nabla u &= \nu\overline{\sigma_f}u \quad \text{in } \Omega,\\ u &= 0 \qquad \text{on } \partial\Omega.\end{aligned} \tag{10}$$

In the convergence of the eigenvectors, $O(\varepsilon)$ *is to be understood in the sense of the* $L^2(\Omega\times V)$*-norm. The positive definite homogeneous tensor* $D=(D_{ij})_{1\le i,j\le d}$ *and the*

homogeneous fission cross section $\overline{\sigma_f}$ *in* (10) *are defined by*

$$
\begin{aligned}
D_{ij} &= \int_V \int_Y \frac{(\psi_\infty^+(y,\mu))^2}{\Sigma(y)} \left(\mu_i + \mu\cdot\nabla\theta^i(y,\mu)\right) \mu_j d\mu\, dy, \\
\overline{\sigma_f} &= \int_Y \sigma_f(y) \left(\int_V \psi_\infty^+(y,\mu) d\mu\right)^2 dy,
\end{aligned} \tag{11}
$$

where $(\mu_i)_{1\le i\le d}$ *are the coordinates of* μ *in* $\mathbb{R}^d$ *and* $(\theta^i)_{1\le i\le d}$ *are the zero mean solutions of*

$$
\begin{aligned}
&-\mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\theta^i + Q\theta^i = \mu_i \left(\mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\right), \\
&y \mapsto \theta^i(y,\mu) \ \textit{is } Y\textit{-periodic}
\end{aligned} \tag{12}
$$

with the local scattering operator being defined by

$$
\begin{aligned}
Qw &= \Sigma^\infty \left(\psi_\infty^+ w \int_V \psi_\infty^+(x,\mu')d\mu' - \psi_\infty^+ \int_V w(x,\mu')\psi_\infty^+(x,\mu')d\mu'\right), \\
\Sigma^\infty &= \Sigma_s + \lambda_\infty \sigma_f.
\end{aligned} \tag{13}
$$

Due to the even parity formulation, the homogeneous coefficients given in this theorem are nonstandard. However, they are similar to those obtained in the homogenization of heterogeneous diffusion problems [29], and it can be shown [8] that they correspond to those derived in the setting of first-order transport [5, 26, 27] . Notice, however, that they slightly differ from the coefficients derived physically in [12].

REMARK 3.3. *Hypothesis* (H6) *is not optimal. We do not need it to prove the convergence of the eigenvalues and eigenvectors, i.e., to replace the rates of convergence* $O(\varepsilon)$ *and* $O(\varepsilon^3)$ *in* (9) *by* $o(1)$ *and* $o(\varepsilon^2)$*, respectively. Hypothesis* (H5) *can also be considerably weakened. We obtained in* [5] *the convergence of the eigenelements for cross sections that are only uniformly bounded from above and below by positive constants.*

In practical nuclear reactor computations, the number of assemblies, or equivalently ε^{-1}, is not very large, and the expansion given in the previous theorem is not sufficiently accurate. The aim of the next theorem is to give a third-order corrector for the smallest eigenvalue λ_1^ε, the only one of physical interest. Numerical simulations also show a better accuracy for the associated positive eigenvector [9], even if a theoretical proof is given only for source problems.

First-order correctors have already been addressed in homogeneous transport theory [33] as well as in heterogeneous diffusion theory [3, 32]. The results obtained in these works show that the first-order corrector for the eigenvalues can strongly depend on the geometry. In transport theory, the first-order corrector is driven by the neutron leakage at the boundary of the core. Since the neutron mean free path is comparable to the size of the unit cell, the boundary $\partial\Omega$ has a direct influence on the leakage. We need to define our geometry more precisely in order to obtain an asymptotic neutron leakage. We study here the case of a cubic domain with symmetric cells as follows.

HYPOTHESIS 3.4.

(H7) Ω *is the unit cube* $(0,1)^d$*. It is composed of* N^d *identical cells, where* $N = \frac{1}{\varepsilon} \in \mathbb{N}$.

(H8) *Hypotheses* (H5) *and* (H6) *are satisfied.*

(H9) *The periodicity cell* Y *is symmetric in the following sense. The cross sections* Σ*,* Σ_s*, and* σ_f *are symmetric with respect to the (hyper)planes parallel to the*

sides of Y and splitting Y in two identical parts. Furthermore these cross sections are invariant by the rotations that preserve Y (rotations of angle $\pi/2$).

Then we have the following first-order corrector for the smallest eigenvalue.

THEOREM 3.5. *Assume that Hypothesis* 3.4 *is satisfied. Then there exists a constant extrapolation length L such that the first eigenvalue ω_1^ε of the elliptic problem*

$$\begin{aligned} -\nabla D\nabla\Phi^\varepsilon &= \omega^\varepsilon\overline{\sigma_f}\Phi^\varepsilon \text{ in } \Omega, \\ \Phi^\varepsilon + \varepsilon L\frac{\partial\Phi^\varepsilon}{\partial n} &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{14}$$

satisfies

$$\lambda_1^\varepsilon = \lambda_\infty + \varepsilon^2\omega_1^\varepsilon + O(\varepsilon^{7/2}). \tag{15}$$

REMARK 3.6. *To our knowledge, this extrapolation length can unfortunately not be computed explicitly. In the case of homogeneous cells, the value of L can be derived from Chandrasekhar's H function* [18]. *An approximate value is $L_0 = 0.7104$ (see, e.g.,* [20, *Chapter* 21]*). In general, L is defined as the limit when the first coordinate $x_1 \to \infty$ of the solution to a conservative transport problem posed on the half space $x_1 > 0$ with some suitable boundary conditions. The analysis of this problem is part of our section* 6. *Since the neutron leakage is positive, we believe that L is always positive, although we do not have a mathematical proof for it. In case $L < 0$, which is physically unrealistic,* (14) *is well posed for ε small enough only.*

REMARK 3.7. *Up to some slight modifications in the proof, the theorem can be extended to the case of cells satisfying all symmetries stated in* (H9) *but the invariance by rotation. Then L is constant on each side of $\partial\Omega$, but not necessarily on the whole $\partial\Omega$.*

We now give a proof for Theorems 3.2 and 3.5. They will rely on some results interesting in themselves for related source problems. We present them in the remaining part of this section and postpone the proofs to the following sections. The proof of Theorem 3.2 relies on the analysis of an equivalent eigenvalue problem for the factored function u_ε defined by

$$u_\varepsilon = \frac{\varphi^{\varepsilon+}}{\psi_\varepsilon^+}, \tag{16}$$

where $\varphi^{\varepsilon+}$ is a solution of (4). Since ψ_∞^+ is positive and regular by virtue of Theorem 2.3, $\varphi^{\varepsilon+} \mapsto u_\varepsilon$ is uniquely defined by (16). The derivation of a transport equation for u_ε uses the following identity:

$$\mu\cdot\nabla\frac{1}{\Sigma}\mu\cdot\nabla(u\psi) = u\mu\cdot\nabla\frac{1}{\Sigma}\mu\cdot\nabla\psi + \frac{1}{\psi}\mu\cdot\nabla\frac{\psi^2}{\Sigma}\mu\cdot\nabla u. \tag{17}$$

An analogous relation was first used in the homogenization of heterogeneous diffusion eigenvalue problems [29]. Plugging (16) into (4) and using (6) and (17), we obtain

$$\begin{aligned} &-\varepsilon^2\mu\cdot\nabla\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}\mu\cdot\nabla u_\varepsilon + \Sigma_s^\varepsilon\left(\psi_\varepsilon^+ u_\varepsilon\int_V \psi_\varepsilon^+(x,\mu')d\mu' - \psi_\varepsilon^+\int_V u_\varepsilon(x,\mu')\psi_\varepsilon^+(x,\mu')d\mu'\right) \\ &\quad = \lambda^\varepsilon\sigma_f^\varepsilon\psi_\varepsilon^+\int_V u_\varepsilon(x,\mu')\psi_\varepsilon^+(x,\mu')d\mu' - \lambda_\infty\sigma_f^\varepsilon u_\varepsilon\psi_\varepsilon^+\int_V \psi_\varepsilon^+(x,\mu')d\mu' \quad \text{in } \Omega\times V. \end{aligned} \tag{18}$$

Let us now derive the boundary conditions satisfied by u_ε. We deduce from (4) and (16) that

$$u_\varepsilon(\psi_\varepsilon^+ - \frac{\varepsilon}{\Sigma^\varepsilon}\mu\cdot\nabla\psi_\varepsilon^+) - \frac{\varepsilon}{\Sigma^\varepsilon}\psi_\varepsilon^+\mu\cdot\nabla u_\varepsilon \quad = \quad 0 \qquad \text{on } \Gamma_-.$$

Introduce $\psi_\varepsilon = \psi_\varepsilon^+ - \frac{\varepsilon}{\Sigma^\varepsilon}\mu\cdot\nabla\psi_\varepsilon^+$. From the equivalence presented in Appendix A, ψ_ε is the solution of the first-order criticality eigenvalue problem

$$\frac{1}{\varepsilon}\mu\cdot\nabla\psi_\varepsilon + \frac{1}{\varepsilon^2}\left(\Sigma^\varepsilon\psi_\varepsilon - \Sigma_\infty^\varepsilon\psi_\varepsilon\right) = 0, \tag{19}$$

where $\Sigma_\varepsilon^\infty = \Sigma_s^\varepsilon + \lambda_\infty\sigma_f^\varepsilon$. It is positive (see [5] and [8, Theorem II.2.1.1]). Thus we obtain the boundary conditions for u_ε:

$$u_\varepsilon - \frac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla u_\varepsilon = 0 \qquad \text{on } \Gamma_-\,. \tag{20}$$

This enables us to recast (18) as

$$\begin{array}{rcll} A_\varepsilon u_\varepsilon & = & \nu^\varepsilon F_\varepsilon u_\varepsilon & \text{in } \Omega\times V, \\ u_\varepsilon - \dfrac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla u_\varepsilon & = & 0 & \text{on } \Gamma_-, \end{array} \tag{21}$$

where we have defined

$$A_\varepsilon u = -\mu\cdot\nabla\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}\mu\cdot\nabla u + \frac{1}{\varepsilon^2}Q_\varepsilon u, \tag{22}$$

$$\nu^\varepsilon = \frac{\lambda^\varepsilon - \lambda_\infty}{\varepsilon^2}, \tag{23}$$

$$Q_\varepsilon u = \Sigma_\varepsilon^\infty\left(\psi_\varepsilon^+ u\int_V \psi_\varepsilon^+(x,\mu')d\mu' - \psi_\varepsilon^+\int_V u(x,\mu')\psi_\varepsilon^+(x,\mu')d\mu'\right), \tag{24}$$

$$F_\varepsilon u = \psi_\varepsilon^+\sigma_f^\varepsilon\int_V \psi_\varepsilon^+(x,\mu')\,u(x,\mu')d\mu'. \tag{25}$$

Following the results given in section 2, ψ_ε^+ is smooth and uniformly positive. Thus problems (4) and (21) are equivalent in $W^2(\Omega\times V)$. We write (21) as

$$\begin{array}{rcl} \dfrac{1}{\nu^\varepsilon}u_\varepsilon & = & S_\varepsilon u_\varepsilon, \\ S_\varepsilon & = & A_\varepsilon^{-1}F_\varepsilon. \end{array} \tag{26}$$

We will see in Theorem 3.9 that $S_\varepsilon \in \mathcal{L}(L^2(\Omega\times V))$. The first step in the asymptotic analysis of problem (26), or equivalently (4), consists in studying the following source problem:

$$\begin{array}{ll} A_\varepsilon w_\varepsilon = F_\varepsilon q & \text{in } \Omega\times V, \\ w_\varepsilon - \dfrac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla w_\varepsilon = 0 & \text{on } \Gamma_-, \end{array} \tag{27}$$

where $q(x,\mu) \in L^2(\Omega\times V)$ is a given source term satisfying $q(x,\mu) = q(x,-\mu)$. Let us recall that we are interested only in solutions of the form $w_\varepsilon(x,-\mu) = w_\varepsilon(x,\mu)$. We first state an a priori estimate for the solutions of source problems.

LEMMA 3.8. *Let $q(x,\mu) \in L^2(\Omega \times V)$ and $g \in L^2(\Gamma_-, d\xi)$. Assume that w_ε is a solution to*

$$\begin{aligned} &A_\varepsilon w_\varepsilon = F_\varepsilon q \quad \text{in } \Omega \times V, \\ &w_\varepsilon - \frac{\varepsilon \psi_\varepsilon^+}{\psi_\varepsilon \Sigma^\varepsilon} \mu \cdot \nabla w_\varepsilon = g \quad \text{on } \Gamma_-. \end{aligned} \tag{28}$$

Then we have

$$\begin{aligned} &\|\mu \cdot \nabla w_\varepsilon\| + \|w_\varepsilon\| + \frac{1}{\sqrt{\varepsilon}} \|w_\varepsilon\|_{L^2(\partial\Omega \times V, d\xi)} + \frac{1}{\varepsilon} \|w_\varepsilon - \int_V w_\varepsilon\| \\ \leq \;& C\|q\| + \frac{C}{\sqrt{\varepsilon}} \|g\|_{L^2(\Gamma_-, d\xi)}, \end{aligned} \tag{29}$$

where $\|\cdot\|$ is the $L^2(\Omega \times V)$-norm and where the measure $d\xi$ on $\partial\Omega \times V$ is given by $d\xi = |\mu \cdot n(x)| d\mu d\sigma$, with $d\sigma$ the surface measure on $\partial\Omega$.

The proof of this lemma is given in section 4. This energy estimate is a key result in our analysis. It enhances the interest of the variational formulation associated with the even parity flux formulation. The asymptotic behavior of the source problem is stated in the following result.

THEOREM 3.9. *Assume that Hypothesis 3.1 is satisfied. Let $q \in L^2(\Omega \times V)$ be a given source term. Then problem (27) admits a unique solution w_ε, which converges strongly to $w \in H_0^1(\Omega)$ in $L^2(\Omega \times V)$ as $\varepsilon \to 0$, where w is the solution of the diffusion problem*

$$\begin{aligned} -\nabla D \nabla w &= \overline{q} \quad \text{in } \Omega, \\ w &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{30}$$

Here we have

$$\overline{q}(x) = \int_Y \sigma_f(y) \left\{ \int_V \psi_\infty^+(y,\mu') d\mu' \int_V \psi_\infty^+(y,\mu) q(x,\mu) d\mu \right\} dy.$$

Moreover, assume that there exists $M \in \mathbb{N}$ such that $q(x,\mu) = \sum_{m=1}^M q_m(x) h_m(\mu)$, where $q_m \in C^{2,\alpha}(\Omega)$ and $h_m \in L^2(V)$, $1 \leq m \leq M$. Then we have the error estimate in $L^2(\Omega \times V)$:

$$w_\varepsilon - w = O(\varepsilon).$$

The proof of the well-posedness of (27) is given in section 4. The proof of the asymptotic behavior stated in this theorem is postponed to section 5. We deduce from this theorem the pointwise convergence of S_ε to the homogenized operator $S \in \mathcal{L}(L^2(\Omega \times V))$ defined by

$$q \mapsto Sq = w,$$

where w is the solution to (30). The *compact convergence* of the sequence S_ε (in the sense given in Appendix B) is asserted by the following lemma.

LEMMA 3.10. *Let x_ε be a sequence of elements in the unit ball of $L^2(\Omega \times V)$. Then $S_\varepsilon x_\varepsilon$ is relatively compact.*

Proof. We deduce from the a priori estimate (29) of Lemma 3.8 that

$$\|\mu \cdot \nabla (S_\varepsilon x_\varepsilon)\|^2 + \|S_\varepsilon x_\varepsilon\| + \|S_\varepsilon x_\varepsilon\|^2_{L^2(\partial\Omega \times V, d\xi)} \leq C.$$

Then the averaging lemma given in [25] yields that $\int_V S_\varepsilon x_\varepsilon \, d\mu$ is relatively compact in $L^2(\Omega)$. Using (29) once again, we have

$$\|S_\varepsilon x_\varepsilon - \int_V S_\varepsilon x_\varepsilon \, d\mu\| \quad \leq \quad C\varepsilon,$$

which asserts that $S_\varepsilon x_\varepsilon$ is also relatively compact. This completes the proof of this lemma. □

We deduce from Theorem 3.9 and Lemma 3.10 that the sequence of operators S_ε converges compactly to S. Theorem 3.2 is then a straight consequence of Theorem B.1. Since the eigenvectors of the diffusion eigenvalue problem are regular, the error estimates (9) given in Theorem 3.2 are easily derived from Theorem B.2.

In order to prove Theorem 3.5, we need to characterize first-order correctors for the source problem (27). The source term in the criticality eigenvalue problem, coming from the fission term, is regular and isotropic. Therefore, we consider only source terms $q = q(x) \in C^{3,\alpha}(\Omega)$ with $\alpha > 0$.

THEOREM 3.11. *Let w_ε be the solution to* (28) *with $g = 0$ and $(\theta^i)_{1\leq i\leq d}$ defined by* (12). *Assume that Hypothesis* 3.4 *is satisfied and that $q = q(x) \in C^{3,\alpha}(\Omega)$. Then there exists a constant L, independent of q, such we have the following result.*

Let w_0 and w_{10} be the solutions of

$$\left\{ \begin{array}{rcll} -\nabla D \nabla w_0 & = & q & \text{in } \Omega, \\ w_0 & = & 0 & \text{on } \partial\Omega, \end{array} \right. \tag{31}$$

$$\left\{ \begin{array}{rcll} -\nabla D \nabla w_{10} & = & 0 & \text{in } \Omega, \\ w_{10} + L\dfrac{\partial w_0}{\partial n} & = & 0 & \text{on } \partial\Omega. \end{array} \right. \tag{32}$$

Then, denoting by $\theta^i_\varepsilon(x,\mu) = \theta^i(\frac{x}{\varepsilon},\mu)$, we have

$$\left\| w_\varepsilon - \left[w_0 + \varepsilon \left(\theta^i_\varepsilon \frac{\partial w_0}{\partial x_i} + w_{10} \right) \right] \right\|_{L^2(\Omega\times V)} = O(\varepsilon^{3/2}). \tag{33}$$

This theorem is proven in section 5. Let us focus on the proof of Theorem 3.5. We obtained in Theorem 3.2 that $\lambda^\varepsilon = \lambda_\infty + \varepsilon^2 \nu^\varepsilon$, where the lth eigenvalue ν^ε_l converges to ν_l as $\varepsilon \to 0$. Now we are interested in the limit of the corrector for the smallest eigenvalue $\xi^\varepsilon_1 = \frac{\nu^\varepsilon_1 - \nu_1}{\varepsilon}$. Denote by s^ε the solution of

$$A_\varepsilon s^\varepsilon = \nu_1 F_\varepsilon u_1.$$

This problem admits a unique solution, as seen in Theorem 3.9. Multiplying this equation by u_ε, the solution of $A_\varepsilon u_\varepsilon = \nu^\varepsilon_1 F_\varepsilon u_\varepsilon$, and integrating over $\Omega \times V$, we obtain, since A_ε is self-adjoint, that

$$\begin{aligned} \nu^\varepsilon_1 (s^\varepsilon, F_\varepsilon u_\varepsilon) &= \nu_1 (u_1, F_\varepsilon u_\varepsilon), \\ (\nu^\varepsilon_1 - \nu_1)(u_1, F_\varepsilon u_\varepsilon) &= \nu^\varepsilon_1 (u_1 - s^\varepsilon, F_\varepsilon u_\varepsilon), \\ \left(\frac{\nu^\varepsilon_1 - \nu_1}{\varepsilon} \right) (u_1, F_\varepsilon u_\varepsilon) &= \nu^\varepsilon_1 \left(\frac{u_1 - s^\varepsilon}{\varepsilon}, F_\varepsilon u_\varepsilon \right). \end{aligned} \tag{34}$$

Since u_1 is sufficiently smooth, we deduce from Theorem 3.11 the following expansion in $L^2(\Omega \times V)$:

$$s^\varepsilon = u_1 + \varepsilon \theta^i_\varepsilon \frac{\partial u_1}{\partial x_i} + \varepsilon w + O(\varepsilon^{3/2}), \tag{35}$$

where w is the solution to (32) with w_0 replaced by u_1. Then

$$\left(\frac{s^\varepsilon - u_1}{\varepsilon}, F_\varepsilon u_1\right) \quad = \quad (w, F_\varepsilon u_1) + \left(F_\varepsilon \theta^i_\varepsilon \frac{\partial u_1}{\partial x_i}, u_1\right) + O(\varepsilon).$$

The first term on the right-hand side clearly converges to $\overline{\sigma}_f(u_1, w)$. The second term is given by

$$\int_\Omega \left[\int_V \theta^i\left(\frac{x}{\varepsilon}, \mu\right) \sigma_f\left(\frac{x}{\varepsilon}\right) \psi^+_\infty\left(\frac{x}{\varepsilon}, \mu\right) d\mu \int_V \psi^+_\infty\left(\frac{x}{\varepsilon}, \mu'\right) d\mu'\right] \frac{\partial u_1}{\partial x_i}(x) u_1(x) dx.$$

From the symmetry properties of Y given in Hypothesis 3.4, we easily deduce (see also Lemma 5.1 in section 5) that

$$\int_Y \int_V \theta^i(y, \mu) \sigma_f(y) \psi^+_\infty(y, \mu) d\mu \int_V \psi^+_\infty(y, \mu') d\mu' dy = 0.$$

Since u_1 is regular, a Taylor expansion of u_1 yields that $(F_\varepsilon \theta^i_\varepsilon \frac{\partial u_1}{\partial x_i}, u_1) = O(\varepsilon)$. Therefore, since $u_\varepsilon = u_1 + O(\varepsilon)$, the right-hand side of (34) is equal to $-\nu_1 \overline{\sigma}_f(u_1, w) + O(\varepsilon^{1/2})$. The left-hand side of (34) is equal to $\xi^\varepsilon_1 \overline{\sigma}_f + O(\varepsilon)$. The first-order corrector for the eigenvalue takes the form

$$\xi^\varepsilon_1 \; = \; -\nu_1(u_1, w) + O(\varepsilon^{1/2}).$$

At last we deduce from [20, Proposition 3, section 5, Chapter 21], or from elementary computations on the eigenvalues of a homogeneous diffusion problem in a cube, that the smallest eigenvalue of

$$\begin{array}{rcll} -\nabla D \nabla \Phi^\varepsilon & = & \omega^\varepsilon \overline{\sigma_f} \Phi^\varepsilon & \text{in } \Omega, \\ \Phi^\varepsilon + \varepsilon L \dfrac{\partial \Phi^\varepsilon}{\partial n} & = & 0 & \text{on } \partial\Omega, \end{array} \tag{36}$$

satisfies u_1 and Φ^ε being normalized in $L^2(\Omega)$,

$$\omega^\varepsilon_1 = \nu_1 - \varepsilon \nu_1(u_1, w) + O(\varepsilon^{3/2}). \tag{37}$$

This concludes the proof of Theorem 3.5.

4. A priori estimates and analysis of the source problem. This section presents some preliminary results on the source problem (27), which will prove useful in the study of its asymptotic behavior as $\varepsilon \to 0$. We give a proof for Lemma 3.8 and for the well-posedness of problem (27) as stated in Theorem 3.9. We will use the notation

$$\langle u \rangle = \int_V u(\mu) d\mu.$$

One of the interesting properties of the even parity flux is that it allows us to use a variational formulation. Introduce the bilinear form $a_\varepsilon(u, v)$

$$\begin{array}{rcl} a_\varepsilon(u, v) & = & \displaystyle\int_\Omega \int_V \frac{(\psi^+_\varepsilon)^2}{\Sigma^\varepsilon} \mu \cdot \nabla u \, \mu \cdot \nabla v \, d\mu \, dx + \frac{1}{\varepsilon} \int_{\partial\Omega} \int_V |\mu \cdot n| \, h_\varepsilon \, uv \, d\mu \, d\sigma \\ & + & \displaystyle\frac{1}{\varepsilon^2} \int_\Omega \int_V (Q_\varepsilon u) \, v \, d\mu \; dx, \end{array} \tag{38}$$

where the function h_ε is defined on $\partial\Omega \times V$ by $h_\varepsilon = \psi_\varepsilon \psi_\varepsilon^+$ on Γ_- and $h_\varepsilon(x,\mu) = h_\varepsilon(x,-\mu)$. According to the results of section 2, h_ε is uniformly positive. An integration by parts shows that (28) is equivalent to finding $w_\varepsilon \in \mathcal{V}$ such that

$$a_\varepsilon(w_\varepsilon, v) = (F_\varepsilon q, v) + \frac{1}{\varepsilon}\int_{\partial\Omega}\int_V |\mu\cdot n| h_\varepsilon\, g\, v\, d\mu\, d\sigma \qquad \forall v \in \mathcal{V}, \tag{39}$$

where $\mathcal{V}$ is the Hilbert space

$$\mathcal{V} = \{v \in W^2(\Omega\times V) \text{ s.t. } v \in L^2(\partial\Omega\times V, d\xi),\ v(x,\mu) = v(x,-\mu)\}. \tag{40}$$

The bilinear form a_ε is bicontinuous in $\mathcal{V}$. We deduce that a_ε is symmetric from the identity $\int_V (Q_\varepsilon u) v\, d\mu = \int_V u(Q_\varepsilon v)\, d\mu$. Now choosing $v = w_\varepsilon$ in (39), we obtain that

$$\begin{aligned} &\|\mu\cdot\nabla w_\varepsilon\|^2 + \frac{1}{\varepsilon}\|w_\varepsilon\|^2_{L^2(\partial\Omega\times V,d\xi)} + \frac{1}{\varepsilon^2}(Q_\varepsilon w_\varepsilon, w_\varepsilon) \\ \le\ & C a_\varepsilon(w_\varepsilon,w_\varepsilon) \le C\|q\|\,\|w_\varepsilon\| + \frac{C}{\varepsilon}\|g\|_{L^2(\partial\Omega\times V,d\xi)}\|w_\varepsilon\|_{L^2(\partial\Omega\times V,d\xi)}. \end{aligned} \tag{41}$$

A Poincaré-like inequality in transport theory (see [5, 8]) yields that

$$\|w_\varepsilon\| \le C\,(\|\mu\cdot\nabla w_\varepsilon\| + \|w_\varepsilon\|_{L^2(\partial\Omega\times V,d\xi)}), \tag{42}$$

where C is a constant independent of w_ε. Therefore, the coercivity of a_ε is easily deduced from the positiveness of the collision operator Q_ε that we prove now.

LEMMA 4.1. *Let $f \in L^\infty(V)$ be a positive function. Then the operator $\mathcal{Q}$ defined by*

$$\mathcal{Q}u(\mu) = f(\mu)u(\mu)\langle f\rangle - f(\mu)\langle f\,u\rangle$$

satisfies the property

$$(\mathcal{Q}u, u) \ge (\inf_V f)^2 \|u - \langle u\rangle\|^2_{L^2(V)},$$

where $(\cdot,\cdot)$ is the usual scalar product of $L^2(V)$.

Proof. Some computations yield

$$\begin{aligned} (\mathcal{Q}u,u) &= \int_V\int_V f(\mu)f(\mu')u(\mu)(u(\mu)-u(\mu'))d\mu d\mu' \\ &= \frac{1}{2}\int_V\int_V f(\mu)f(\mu')(u(\mu)-u(\mu'))^2 d\mu d\mu' \\ &\ge (\inf_V f)^2\frac{1}{2}\int_V\int_V (u(\mu)-u(\mu'))^2 d\mu d\mu' \ =\ (\inf_V f)^2\int_V (u(\mu)-\langle u\rangle)^2\, d\mu. \end{aligned}$$

This concludes the proof of the lemma. ∎

From this lemma and the Poincaré inequality (42), we obtain that

$$\begin{aligned} &\|\mu\cdot\nabla w_\varepsilon\| + \|w_\varepsilon\| + \frac{1}{\sqrt{\varepsilon}}\|w_\varepsilon\|_{L^2(\partial\Omega\times V,d\xi)} + \frac{1}{\varepsilon}\|w_\varepsilon - \int_V w_\varepsilon\| \\ \le\ & C\sqrt{a_\varepsilon(w_\varepsilon,w_\varepsilon)} \le C\left(\|q\| + \frac{1}{\sqrt{\varepsilon}}\|g\|_{L^2(\partial\Omega\times V,d\xi)}\right). \end{aligned} \tag{43}$$

This concludes the proof of Lemma 3.8. Moreover, the existence and uniqueness of a solution to (28) is a straight consequence of the Lax–Milgram theory. This also asserts the well-posedness and boundedness of S_ε in $\mathcal{L}(L^2(\Omega\times V))$, and the first part of Theorem 3.9 is complete. We also need an analogous result for nonhomogeneous boundary conditions. We have the maximum principle as follows.

PROPOSITION 4.2. *Let $g\in L^\infty(\Gamma_-)$. There exists a unique solution to* (28) *with $q=0$ in $L^\infty(\Omega\times V)$. Furthermore, we have that $\|w_\varepsilon\|_{L^\infty(\Omega\times V)}\le\|g\|_{L^\infty(\Gamma_-)}$.*

Proof. We use the maximum principle stated for first-order transport equations in [20, Chapter 21] and the equivalence presented in Appendix A. Let $\alpha>0$ and the sequence of problems

$$\begin{aligned} -\mu\cdot\nabla\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}\mu\cdot\nabla w_\alpha^\varepsilon+\frac{1}{\varepsilon^2}Q_\varepsilon w_\alpha^\varepsilon+\alpha w_\alpha^\varepsilon &= 0 \quad \text{in } \Omega\times V,\\ w_\alpha^\varepsilon-\frac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla w_\alpha^\varepsilon &= g \quad \text{on } \Gamma_-. \end{aligned}$$

We denote $\check f(\mu)=f(-\mu)$ for every function f. We deduce from the results recalled in Appendix A that

$$w_\alpha^\varepsilon=\frac{\psi_\varepsilon\psi_\alpha^\varepsilon+\check\psi_\varepsilon\check\psi_\alpha^\varepsilon}{\psi_\varepsilon+\check\psi_\varepsilon}.$$

Here, ψ_ε is the positive solution of (19) and ψ_α^ε is the solution to

$$\begin{aligned} \frac{1}{\varepsilon}\mu\cdot\nabla\psi_\alpha^\varepsilon+\frac{\Sigma_\varepsilon^\infty}{\varepsilon^2}\left(\frac{1}{\psi_\varepsilon}\psi_\alpha^\varepsilon\int_V\psi_\varepsilon-\frac{1}{\psi_\varepsilon}\int_V\psi_\alpha^\varepsilon\psi_\varepsilon\right)+\alpha\psi_\alpha^\varepsilon &= 0 \quad \text{in } \Omega\times V,\\ \psi_\alpha^\varepsilon &= g \quad \text{on } \Gamma_-. \end{aligned}$$

We deduce from [20, Proposition 7, Chapter 21, section 2] that $\|\psi_\alpha^\varepsilon\|_{L^\infty}\le\|g\|_{L^\infty}$ independently of α. Therefore, $\|w_\alpha^\varepsilon\|_{L^\infty}\le\|g\|_{L^\infty}$ independently of α. Thus there exists a subsequence of w_α^ε that converges to w_ε in $L^\infty(\Omega\times V)$ weak$*$ as $\alpha\to0$. Using standard techniques (see Bardos [10] and Appendix A), we verify that w_ε is a solution of (28) and satisfies $\|w_\varepsilon\|_{L^\infty(\Omega\times V)}\le\|g\|_{L^\infty(\Gamma_-)}$. We know from Lemma 3.8 that

$$\|w_\varepsilon\|\le\frac{C}{\sqrt{\varepsilon}}\|g\|_{L^2(\Gamma_-,d\xi)}.$$

Therefore, the solution to (28) in $L^\infty(\Omega\times V)$ is unique. ∎

Before addressing the asymptotic convergence of the solutions to (27), we need one more result on source problems in infinite media. Let us introduce the factored flux φ^+ defined by $\psi^+=\psi_\infty^+\varphi^+$. Then we easily obtain the following corollary for Theorem 2.5.

COROLLARY 4.3. *Assume that the hypotheses of Theorem* 2.5 *are satisfied. Then problem*

$$\begin{aligned} &-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\varphi^+ + Q\varphi^+ = S \quad \text{in } Y\times V,\\ &y\mapsto\varphi^+(y,\mu) \text{ is } Y\text{-periodic} \end{aligned} \tag{44}$$

admits a unique zero mean solution if and only if

$$\int_Y\int_V S(y,\mu)dy\,d\mu = 0.$$

The same regularity results as in Theorem 2.5 *hold true.*

Since ψ_∞^+ is periodic, we deduce from this corollary that (12) admits a unique zero mean solution. Then the diffusion coefficients are uniquely defined by (11) and do not depend on the mean over $Y \times V$ of the functions θ^i. The following lemma asserts the positive definiteness of the diffusion tensor D and consequently the well-posedness of the diffusion problem (10).

LEMMA 4.4. *The homogeneous tensor D defined by* (11) *is positive definite.*

Proof. Let us multiply (12) by a Y-periodic smooth test function. After integration by parts, we obtain

$$\begin{aligned}&\int_Y\int_V \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\theta^i\mu\cdot\nabla_y v\,d\mu dy + (Q\theta^i,v) - \int_{\partial Y}\int_V \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\theta^i(\mu\cdot n)v\,d\mu d\sigma\\ =\ & -\int_Y\int_V \mu_i\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y v\,d\mu dy + \int_{\partial Y}\int_V \mu_i\frac{(\psi_\infty^+)^2}{\Sigma}v(\mu\cdot n)\,d\mu d\sigma.\end{aligned}$$

The boundary terms in this equation cancel out since $\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\theta^i v$ and $\mu_i\frac{(\psi_\infty^+)^2}{\Sigma}v$ are Y-periodic, whereas the sign of $(\mu\cdot n)$ is reversed from one side of Y to the opposite one. Thus we have

$$\int_Y\int_V \frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_y\theta^i+\mu_i)\mu\cdot\nabla_y v\,d\mu dy + (Q\theta^i,v) = 0$$

for any periodic test function v and in particular for the functions θ^j. We recall that the coefficients D_{ij} are defined by (11). For each vector $\xi\in\mathbb{R}^n$, of components ξ_i, we deduce that (with the convention of summation over the repeated indexes)

$$\begin{aligned}D_{ij}\xi_i\xi_j &= \int_Y\int_V \frac{(\psi_\infty^+)^2}{\Sigma}\mu_j\xi_j(\mu\cdot\nabla_y\theta^i+\mu_i)\xi_i\,d\mu dy\\ &= \int_Y\int_V \frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_y\theta^j+\mu_j)\xi_j(\mu\cdot\nabla_y\theta^i+\mu_i)\xi_i\,d\mu dy\\ &\quad -\int_Y\int_V \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\theta^j\xi_j(\mu\cdot\nabla_y\theta^i+\mu_i)\xi_i\,d\mu dy\\ &= \int_Y\int_V \frac{(\psi_\infty^+)^2}{\Sigma}\left\{\left(\sum_{i=1}^n(\mu\cdot\nabla_y\theta^i+\mu_i)\xi_i\right)^2 + \left(Q(\sum_{i=1}^n\xi_i\theta^i),\sum_{i=1}^n\xi_i\theta^i\right)\right\}d\mu dy.\end{aligned}$$

We deduce from Lemma 4.1 that this expression is nonnegative. Therefore, D is positive. Assume now that it is not definite. Then there exists a vector $\xi\neq 0$ such that $(\mu\cdot\nabla_y\theta^i+\mu_i)\xi_i = 0$ almost everywhere (a.e.) $(y,\mu)\in Y\times V$. Integrate this equality over Y for $\mu\in V$ given. We deduce from the Y-periodicity of θ^i that $\int_Y \mu_i\xi_i = \mu_i\xi_i = 0$ a.e. $\mu\in V$. This yields $\xi = 0$ and a contradiction. Thus the tensor D is positive definite. ∎

5. Analysis of the source problem. This section is devoted to the proof of Theorems 3.9 and 3.11. It is based on the classical method of two-scale asymptotic expansions. Notice that the two-scale convergence introduced in [2] represents a natural framework to prove the convergence stated in Theorem 3.9 [5]. It allows for very mild regularity assumptions on the physical data; however, it does not enable us to obtain the rates of convergence stated in section 3.

5.1. Asymptotic convergence of the source problem. The pointwise convergence of the operator S_ε is stated in Theorem 3.9. We now give a proof of this result.

Proof. It is based on a two-scale analysis of the neutron density. Let us define the ansatz

$$w_\varepsilon(x,\mu) = w_0\left(x,\frac{x}{\varepsilon},\mu\right) + \varepsilon w_1\left(x,\frac{x}{\varepsilon},\mu\right) + \varepsilon^2 w_2\left(x,\frac{x}{\varepsilon},\mu\right) + \zeta^\varepsilon(x,\mu), \tag{45}$$

where the functions $y \mapsto w_k(x,y,\mu)$ are Y-periodic for $0 \le k \le 2$. We derive the equations that $w_k(x,y,\mu)$ must verify to justify this asymptotic expansion. This provides us with an explicit choice for $w_k(x,y,\mu)$. Next, we prove that ζ^ε defined by (45) is of order $O(\varepsilon)$ in $L^2(\Omega\times V)$. This is done first for a source term $q(x,\mu) = \tilde q(x)h(\mu)$ and then extended to the case $q \in L^2(\Omega\times V)$ by density.

(i) We remark that the differentiation operator is now given by $\mu\nabla = \mu\nabla_x + \frac{1}{\varepsilon}\mu\nabla_y$. Inserting (45) into (27) and neglecting ζ_ε, we obtain

$$\begin{aligned} &- \left(\mu\cdot\nabla_x + \frac{1}{\varepsilon}\mu\cdot\nabla_y\right)\frac{(\psi_\infty^+)^2}{\Sigma}\left(\mu\cdot\nabla_x + \frac{1}{\varepsilon}\mu\cdot\nabla_y\right)(w_0+\varepsilon w_1+\varepsilon^2 w_2)\\ &+ \frac{1}{\varepsilon^2}Q(w_0+\varepsilon w_1+\varepsilon^2 w_2) = \sigma_f\psi_\infty^+\int_V \psi_\infty^+(x,\mu')q(x,\mu')d\mu'. \end{aligned} \tag{46}$$

The term of order -2 in the expansion in powers of ε yields

$$-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_0 + Qw_0 = 0. \tag{47}$$

Following the same development as in section 3 we obtain that $w_0\psi_\infty^+$ is a solution of (6) for any given x. Then by virtue of Theorem 2.3, we have $w_0 = w_0(x)$. From the boundary conditions for w_ε we deduce that w_0 must vanish on $\partial\Omega$.

Taking into account the form of w_0, we obtain for the terms of order ε^{-1} that

$$-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_1 + Qw_1 = \mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x w_0. \tag{48}$$

This equation was posed on $Y\times V$ for every given $x\in\Omega$ and we deduce from Corollary 4.3 that $w_1(y,\mu) = \theta^i(y,\mu)\frac{\partial w_0}{\partial x_i}(x) + w_{10}(x)$, where θ^i is defined by (12) and $w_{10}(x)$ is still undetermined. We choose here $w_{10}(x) = 0$.

Consider now the zeroth-order equation. We have

$$\begin{aligned} &-\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2 w_0 - \left(\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x + \mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\right)w_1\\ &-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_2 + Qw_2 = \sigma_f\psi_\infty^+\int_V \psi_\infty^+(y,\mu')q(x,\mu')d\mu'. \end{aligned} \tag{49}$$

Following Corollary 4.3, this equation admits a solution when the source term is of zero mean. It implies that w_0 satisfies the equation

$$\begin{aligned} &\int_V\int_Y\left[-\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2 w_0 - \left(\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x + \mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\right)w_1\right]d\mu dy\\ =\ &\int_V\int_Y\left(\sigma_f(y)\psi_\infty^+(y,\mu)\int_V\psi_\infty^+(y,\mu')q(x,\mu')d\mu'\right)d\mu dy. \end{aligned}$$

Replacing w_1 by its expression in terms of w_0, we obtain that w_0 is a solution to (30). Since D is positive definite according to Lemma 4.1, we deduce that w_0 is uniquely defined in $H_0^1(\Omega)$. The expression of w_1 is also known. It remains to define w_2. Equation (49) can be recast as

$$-\mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_2 + Qw_2 = \left(\sigma_f\psi_\infty^+\int_V \psi_\infty^+ q \,-\, \overline{q}\right) + (h_{ij}(y,\mu) - D_{ij})\,\frac{\partial^2 w_0}{\partial x_j \partial x_i}$$

owing to the expression of w_1 and (30), where

$$h_{ij} = \frac{(\psi_\infty^+)^2}{\Sigma}\mu_j(\mu_i + \mu\cdot\nabla_y\theta^i) + \mu_j\,\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\theta^i.$$

We define the functions w_{2a} and w_{2ij}, $1\le i,j\le n$, recalling that $q(x,\mu)=\tilde{q}(x)h(\mu)$, by

$$-\mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_{2a} + Qw_{2a} = \sigma_f\psi_\infty^+\int_V \psi_\infty^+ h\,d\mu' \,-\, \int_Y\int_V\left(\sigma_f\psi_\infty^+\int_V\psi_\infty^+ h\,d\mu'\right)d\mu dy$$

$$-\mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_{2ij} + Qw_{2ij} = h_{ij} - D_{ij}.$$

We easily check that the source terms have zero mean. Consequently the functions w_{2a} and w_{2ij} are defined up to an additive constant and are smooth according to Corollary 4.3. By linearity of transport, we obtain that w_2 is equal to

$$w_2(x,y,\mu) = w_{2a}(y,\mu)\tilde{q}(x) + w_{2ij}(y,\mu)\frac{\partial^2 w_0}{\partial x_j\partial x_i}(x) + w_{20}(x),$$

where w_{20} is undetermined. We choose here $w_{20}=0$.

(ii) It remains to derive an equation for ζ^ε and prove that this error term is small. Inserting (45) in (27), we obtain from the explicit expressions of w_0, w_1, and w_2 that

$$\begin{aligned} -\mu\cdot\nabla\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}\mu\cdot\nabla\zeta^\varepsilon + \frac{1}{\varepsilon^2}Q_\varepsilon\zeta^\varepsilon &= \varepsilon\zeta_1^\varepsilon + \varepsilon^2\zeta_2^\varepsilon \quad \text{in } \Omega\times V,\\ \zeta^\varepsilon - \frac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla\zeta^\varepsilon &= \varepsilon\zeta_3^\varepsilon + \varepsilon^2\zeta_4^\varepsilon + \varepsilon^3\zeta_5^\varepsilon \qquad \text{on } \Gamma_-, \end{aligned} \tag{50}$$

where $\zeta_i^\varepsilon(x,\mu) = \zeta_i(x,\frac{x}{\varepsilon},\mu)$ and

$$\begin{aligned} \zeta_1 &= -\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2 w_1 - \mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_2 - \mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x w_2,\\ \zeta_2 &= -\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2 w_2, \qquad \zeta_3 = -w_1 + \frac{\psi_\infty^+}{\psi_\infty\Sigma}(\mu\cdot\nabla_x w_0 + \mu\cdot\nabla_y w_1),\\ \zeta_4 &= -w_2 + \frac{\psi_\infty^+}{\psi_\infty\Sigma}(\mu\cdot\nabla_x w_1 + \mu\cdot\nabla_y w_2), \qquad \zeta_5 = \frac{\psi_\infty^+}{\psi_\infty\Sigma}\mu\cdot\nabla_x w_2. \end{aligned} \tag{51}$$

Assume first that $\tilde{q}$ is of class $C^{2,\alpha}(\Omega)$. Since D is positive definite, we deduce from [24] that w_0 is of class $C^{4,\alpha}(\Omega)$. The terms w_1 and w_2 are the sums of products of functions depending only on x and of functions depending only on (y,μ). The part depending only on x in the expression of w_1: $\frac{\partial w_0}{\partial x_i}$ is of class $C^{3,\alpha}(\Omega)$, and the one in the expression of w_2: $\tilde{q}$ and $\frac{\partial^2 w_0}{\partial x_j x_i}$ is of class $C^{2,\alpha}(\Omega)$. The parts depending on

(y,μ) in terms θ^i, w_{2a}, and w_{2ij} are regular according to Corollary 4.3. Therefore, the terms ζ_i^ε are of class $C^{0,\alpha}(\Omega\times V)$. From the variational formulation of (50) (see Lemma 3.8 and Proposition 4.2), we deduce that $\|\zeta^\varepsilon\|\le C\varepsilon$, where C is independent of ε. Therefore, we have proven that $w_\varepsilon - w = O(\varepsilon)$, provided q is of the form $q(x,\mu)=\sum_{m=1}^M q_m(x)h_m(\mu)$, where all q_m are of class $C^{2,\alpha}(\Omega)$.

The proof of the convergence in the general case follows from two density arguments. Let us first assume that $\tilde q$ belongs to $L^2(\Omega)$. Consider a sequence $\tilde q_i$ of functions of class $C^{2,\alpha}(\Omega)$ converging to $\tilde q$ strongly in $L^2(\Omega)$. Then for any $\eta>0$ and i large enough we have $\|\tilde q_i-\tilde q\|\le C\eta$. Let w_i^ε be the solution of (27). From the linearity of the transport equation and from its variational formulation, we deduce that $\|w_i^\varepsilon - w_\varepsilon\|\le C\eta$ independently of η and ε. On the other hand we have proven that $w_i^\varepsilon\to w_i$, where w_i is defined as the solution of the diffusion solution with a source term equal to $\tilde q_i$ instead of $\tilde q$. We easily check that $\|w_i-w_0\|\le C\eta$, where w_0 is the solution of (30) with source term $\tilde q$, and deduce the convergence of w_ε to w_0 as $\varepsilon\to 0$.

The same argument is used for the general case $q\in L^2(\Omega\times V)$. Every function $q\in L^2(\Omega\times V)$ admits a spectral decomposition in the canonical Schauder basis of $L^2(\Omega\times V)$

$$q(x,\mu) \;=\; \sum_{m=1}^\infty\sum_{p=1}^\infty \alpha_{mp}\tilde q_m(x)h_p(\mu),$$

where the functions $\tilde q_m$ are vectors of an orthogonal basis in $L^2(\Omega)$ and h_p are vectors of an orthogonal basis in $L^2(V)$. We have that $\sum_{m=1}^\infty\sum_{p=1}^\infty \alpha_{mp}^2<\infty$. Thus the sequence

$$q_{MP}=\sum_{m=1}^M\sum_{p=1}^P \alpha_{mp}\tilde q_m(x)h_p(\mu)$$

converges strongly to q in $L^2(\Omega\times V)$. We deduce that $\|w_{MP}^\varepsilon - w_{MP}\|$ converges to 0 independently of ε and that w_{MP} converges strongly to w (using obvious notation) as $M,P\to\infty$. On the other hand for fixed M,P large enough, the linearity of the transport and diffusion equations yields the strong convergence in $L^2(\Omega\times V)$ of w_{MP}^ε to w_{MP} as $\varepsilon\to 0$. ∎

5.2. First-order corrector for the source problem. This subsection is devoted to the derivation of the first-order corrector of the source problem (27) as stated in Theorem 3.11.

First we introduce some notation. To simplify we assume here that the spatial dimension is $d=3$. Let O be the center of Y. We denote by $\mathcal{P}_p$ the plane (O,x_m,x_n), $m,n\in\{i,j,k\}$ spanned by the vectors e_m, e_n and such that $O\in\mathcal{P}_p$. The index $p\in\{i,j,k\}$ is such that $p\ne m$ and $p\ne n$.

We say that two pairs (y,μ) and (y',μ') of $Y\times V$ are symmetric with respect to $\mathcal{P}_p$ if they satisfy

$$d(y,\mathcal{P}_p)=d(y',\mathcal{P}_p),\quad (y'-y)\;\|\;e_p,\quad \mu'=\mu-2(\mu\cdot e_p)e_p.$$

For a pair (y,μ) and (y',μ') of symmetric points we say that a function ψ is symmetric with respect to $\mathcal{P}_p$ if $\psi(y',\mu')=\psi(y,\mu)$, and that it is skew symmetric if $\psi(y',\mu')=-\psi(y,\mu)$. We now state Lemma 5.1.

LEMMA 5.1. *Assume that the cross sections Σ and Σ_s are symmetric in Y with respect to the planes $\mathcal{P}_p$, $p = i, j, k$. Let (y, μ) and (y', μ') be symmetric points with respect to $\mathcal{P}_p$ for $p \in \{i, j, k\}$.*

(i) *Let f be a source term satisfying $\check{f} = f$ and $f(y, \mu) = \pm f(y', \mu')$ a.e. $(y, \mu) \in Y \times V$. Then the solution of*

$$\begin{aligned} &-\mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\varphi + Q\varphi = f \quad \text{in } Y \times V,\\ &y \mapsto \varphi(y,\mu) \text{ is } Y\text{-periodic} \end{aligned}$$

satisfies the following relation of symmetry $\varphi(y, \mu) = \pm\varphi(y', \mu')$.

(ii) *Let f be a function satisfying that $f(y, \mu) = \pm f(y', \mu')$ a.e. $(y, \mu) \in Y \times V$. Then we have*

$$\mu'\cdot\nabla_{y'} f(y', \mu') = \mu\cdot\nabla_y f(y, \mu).$$

Proof. (i) Define $h(y', \mu') = \varphi(y, \mu)$. We check that $\mu'\cdot\nabla_{y'} h(y', \mu') = \mu\cdot\nabla_y\varphi(y, \mu)$. Since Y is symmetric, the solution ψ_∞^+ and the cross sections are also symmetric. Then

$$-\mu'\cdot\nabla_{y'} \frac{(\psi_\infty^+)^2}{\Sigma}(y', \mu')\mu'\cdot\nabla_{y'} h(y', \mu') = \mu\cdot\nabla_y \frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\varphi$$

and

$$(Qh)(y', \mu') = (Q\varphi)(y, \mu).$$

Therefore, h is solution of the equation

$$-\mu'\cdot\nabla_{y'} \frac{(\psi_\infty^+)^2}{\Sigma}(y', \mu')\mu'\cdot\nabla_{y'} h(y', \mu') + (Qh)(y', \mu') = \pm f(y', \mu').$$

We deduce from the uniqueness of the solution for this equation that $h(y', \mu') = \pm\varphi(y, \mu)$. It concludes the first part of the proof.

(ii) Consider the plane of symmetry $\mathcal{P}_k$. We have $\mu\cdot\nabla_y f = \mu_i \frac{\partial f}{\partial y_i}$. If $i = k$, then $\frac{\partial f}{\partial y_i}(y', \mu') = -\frac{\partial f}{\partial y_i}(y, \mu)$ and $(\mu_i)' = \mu_i' = -\mu_i$. If $i \neq k$, then $\frac{\partial f}{\partial y_i}(y', \mu') = \frac{\partial f}{\partial y_i}(y, \mu)$ and $(\mu_i)' = \mu_i$. In any case we have $(\mu_i \frac{\partial f}{\partial y_i})(y', \mu') = (\mu_i \frac{\partial f}{\partial y_i})(y, \mu)$, and the proof is complete. ∎

We are now in a position to prove Theorem 3.11.

Proof of Theorem 3.11. The first part of the proof is very similar to that of Theorem 3.9. We assume the following ansatz on w_ε:

$$w_\varepsilon(x, \mu) = w_0\left(x, \frac{x}{\varepsilon}, \mu\right) + \varepsilon w_1\left(x, \frac{x}{\varepsilon}, \mu\right) + \varepsilon^2 w_2\left(x, \frac{x}{\varepsilon}, \mu\right) + \varepsilon^3 w_3\left(x, \frac{x}{\varepsilon}, \mu\right) + \zeta^\varepsilon(x, \mu), \tag{52}$$

where the functions $y \mapsto w_i(x, y, \mu)$ are Y-periodic for $0 \leq i \leq 3$. Our objective is to derive some conditions on the terms w_i such that ζ^ε be of order $O(\varepsilon^{3/2})$. The main difference with the proof of Theorem 3.9 is that the asymptotic expansion does not provide suitable boundary conditions for the functions w_i. This difficulty will be overcome by the analysis of a boundary layer problem, which allows us to find boundary conditions for the terms w_i such that the error ζ^ε be of order ε in the vicinity of the boundary but of order $\varepsilon^{3/2}$ globally in $L^2(\Omega \times V)$.

(i) Plugging (52) into (27) and neglecting ζ^ε, we obtain

$$
\begin{aligned}
&-\left(\mu\cdot\nabla_x+\frac{1}{\varepsilon}\mu\cdot\nabla_y\right)\frac{(\psi_\infty^+)^2}{\Sigma}\left(\mu\cdot\nabla_x+\frac{1}{\varepsilon}\mu\cdot\nabla_y\right)(w_0+\varepsilon w_1+\varepsilon^2w_2+\varepsilon^3w_3)\\
&+\frac{1}{\varepsilon^2}Q_\varepsilon(w_0+\varepsilon w_1+\varepsilon^2w_2+\varepsilon^3w_3)=\sigma_f\psi_\infty^+\int_V\psi_\infty^+(y,\mu')q(x)d\mu'.
\end{aligned}
\tag{53}
$$

From the terms of order ε^{-2}, we have $w_0=w_0(x)$, and from the terms of order ε^{-1}, $w_1=\theta^i\frac{\partial w_0}{\partial x_i}+w_{10}$. Here we cannot choose $w_{10}=0$ since it is of order ε. The term of order ε^0 is given by

$$
\begin{aligned}
&-\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2w_0-\left(\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x+\mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\right)w_1\\
&-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_yw_2+Qw_2=\sigma_f\psi_\infty^+\int_V\psi_\infty^+(y,\mu')q(x)d\mu'.
\end{aligned}
$$

This equation admits a solution when (31) is satisfied. We recast the equation for w_2 as follows:

$$
\begin{aligned}
&-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_yw_2+Qw_2\\
=\;&\left(\sigma_f\psi_\infty^+\int_V\psi_\infty^+q-\overline{q}\right)+(h_{ij}(y,\mu)-D_{ij})\frac{\partial^2w_0}{\partial x_j\partial x_i}+\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_xw_{10},
\end{aligned}
$$

where $h_{ij}=\frac{(\psi_\infty^+)^2}{\Sigma}\mu_j(\mu_i+\mu\cdot\nabla_y\theta^i)+\mu_j\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\theta^i$. Since $q=q(x)$, we have

$$
\begin{aligned}
&\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_yw_{2a}+Qw_{2a}=\sigma_f\psi_\infty^+\int_V\psi_\infty^+\,d\mu'-\int_Y\int_V\left(\sigma_f\psi_\infty^+\int_V\psi_\infty^+\,d\mu'\right)d\mu dy\\
&-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_yw_{2ij}+Qw_{2ij}=h_{ij}-D_{ij}.
\end{aligned}
$$

In these equations the source term has zero mean. Thus w_2 can be written as

$$
w_2(x,y,\mu)=w_{2a}(y,\mu)q(x)+w_{2ij}(y,\mu)\frac{\partial^2w_0}{\partial x_j\partial x_i}(x)+\theta^i\frac{\partial w_{10}}{\partial x_i}+w_{20}(x).
$$

Because we are not interested in the terms of order ε^2 we can choose $w_{20}=0$. Let us go one step further in the expansion in powers of ε. The term of order ε yields

$$
\begin{aligned}
&-\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2w_1-\left(\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x+\mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\right)w_2\\
&-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_yw_3+Qw_3=0.
\end{aligned}
$$

This equation admits a solution if and only if the source term

$$
\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2w_1+\left(\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x+\mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y\right)w_2
$$

has zero mean. Replacing the functions w_i by their expressions in terms of w_0, w_{10}, and q, and denoting for simplicity by $\partial_i=\frac{\partial}{\partial x_i}$, we rewrite this compatibility condition

as

$$\left(\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}\left[(\mu_k\mu_j\theta^i)+\mu\cdot\nabla_y(\mu_k w_{2ij})\right]d\mu dy\right)\partial^3_{kji}w_0$$
$$+\left(\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}\mu_i\mu\cdot\nabla_y w_{2a}d\mu dy\right)\partial_i q$$
$$+\left(\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}\left[(\mu\cdot\nabla_x)^2 w_{10}+\mu\cdot\nabla_y\theta^i\mu\cdot\nabla_x\partial_i w_{10}\right]d\mu dy\right)=0.$$

The last term of the left side is given by

$$\left(\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}\left[(\mu\cdot\nabla_x)^2 w_{10}+\mu\cdot\nabla_y\theta^i\mu\cdot\nabla_x\partial_i w_{10}\right]d\mu dy\right)=\nabla D\nabla w_{10}.$$

Using the hypotheses of symmetry on Y, we prove now that the other terms in this expression vanish. We use results stated in Lemma 5.1. First, we easily obtain that

$$\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}(\mu_k\mu_j\theta^i)d\mu dy=0.$$

Indeed θ^i and μ_j are skew symmetric with respect to the plane $\mathcal{P}_i$ and symmetric with respect to the other planes $\mathcal{P}_k$ for $k\neq i$. The product of three skew symmetric functions being skew symmetric at least with respect to one hyperplane, we deduce from the symmetries of $\frac{(\psi_\infty^+)^2}{\Sigma}$ that the integral vanishes. Consider now the term

$$\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}\mu_k\mu\cdot\nabla_y(w_{2ij})d\mu dy.$$

We easily check by symmetry that $D_{ij}=0$ if $i\neq j$. Thus the source term in the equation for w_{2ij} equals h_{ij}. Since θ^i is skew symmetric with respect to $\mathcal{P}_i$,

$$\left(\mu_i+\mu\cdot\nabla_y\theta^i+\mu\cdot\nabla\left(\frac{(\psi_\infty^+)^2}{\Sigma}\theta^i\right)\right)$$

is also skew symmetric. Then w_{2ij} and therefore $\mu\cdot\nabla_y w_{2ij}$ are skew symmetric with respect to $\mathcal{P}_i$. Here again, for all values of k, the integral vanishes. The last term is

$$\int_V\int_Y \frac{(\psi_\infty^+)^2}{\Sigma}\mu_i\mu\cdot\nabla_y w_{2a}d\mu dy.$$

The source terms in the equation for w_{2a} and then w_{2a} and $\mu\cdot\nabla_y w_{2a}$ are symmetric. Since μ_i is skew symmetric with respect to $\mathcal{P}_i$, the corresponding integral vanishes. It follows that w_{10} satisfies (32) on $\Omega\times V$.

We do not give here the equation satisfied by w_3 explicitly. However, w_3, like w_2, is the sum of products of functions of (y,μ) and of functions of x. The terms depending on (y,μ) are regular by hypothesis, and the terms depending on x depend on fifth-order derivatives of w_0, first-order derivatives of q, and second-order derivatives of w_{10}. By hypothesis for q, w_0 and w_{10} are sufficiently regular. This yields that w_3 is well defined and has continuous second-order derivatives. Again w_3 is defined up to an additive function $w_{30}(x)$. We choose $w_{30}=0$.

(ii) It remains to define a constant L such that the term ζ^ε are small. Assume that L is given and insert the expressions of the functions w_i given by (52) into (27). We obtain

$$\begin{aligned} -\mu\cdot\nabla\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}\mu\cdot\nabla\zeta^\varepsilon+\frac{1}{\varepsilon^2}Q_\varepsilon\zeta^\varepsilon &= \varepsilon^2\zeta_1^\varepsilon+\varepsilon^3\zeta_2^\varepsilon \qquad \text{in } \Omega\times V,\\ \zeta^\varepsilon-\frac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla\zeta^\varepsilon &= \varepsilon\zeta_3^\varepsilon+\varepsilon^2\zeta_4^\varepsilon+\varepsilon^3\zeta_5^\varepsilon+\varepsilon^4\zeta_6^\varepsilon \qquad \text{on } \Gamma_-, \end{aligned} \tag{54}$$

where $\zeta_i^\varepsilon(x,\mu)=\zeta_i(x,\frac{x}{\varepsilon},\mu)$ and

$$\begin{aligned} \zeta_1 &= -\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2w_2-\mu\cdot\nabla_x\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y w_3-\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_x w_3,\\ \zeta_2 &= -\frac{(\psi_\infty^+)^2}{\Sigma}(\mu\cdot\nabla_x)^2w_3, \qquad \zeta_3 = -w_1+\frac{\psi_\infty^+}{\psi_\infty\Sigma}(\mu\cdot\nabla_x w_0+\mu\cdot\nabla_y w_1),\\ \zeta_4 &= -w_2+\frac{\psi_\infty^+}{\psi_\infty\Sigma}(\mu\cdot\nabla_x w_1+\mu\cdot\nabla_y w_2),\\ \zeta_5 &= -w_3+\frac{\psi_\infty^+}{\psi_\infty\Sigma}(\mu\cdot\nabla_x w_2+\mu\cdot\nabla_y w_3), \qquad \zeta_6 = \frac{\psi_\infty^+}{\psi_\infty\Sigma}\mu\cdot\nabla_x w_3. \end{aligned} \tag{55}$$

Each term ζ_i gives rise to contributions of order ε^2 except ζ_3. We would like to derive some boundary conditions on w_{10}, which would enable us to cancel out this first-order term. This cannot be done in general because ζ_3 depends on the variables x, y, and μ, whereas w_{10} depends only on x. Replacing w_1 by its expression in terms of w_0 and w_{10}, we find that

$$\zeta_3(x,y,\mu)=\kappa_i(y,\mu)\frac{\partial w_0}{\partial x_i}(x)-w_{10}(x),$$

where $\kappa_i(y,\mu)=(\frac{\psi_\infty^+}{\psi_\infty\Sigma}(\mu_i+\mu\cdot\nabla_y\theta^i)-\theta^i)(y,\mu)$. Let us define the first-order term b^ε as a solution of

$$\begin{aligned} A_\varepsilon b^\varepsilon &= 0 \qquad \text{in } \Omega\times V,\\ b^\varepsilon-\frac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla b^\varepsilon &= \zeta_3^\varepsilon \qquad \text{on } \Gamma_-. \end{aligned} \tag{56}$$

From the linearity of the transport equations, the function b^ε is the sum of $2d$ terms having vanishing boundary conditions on each side of Ω but one. For all $x\in\mathbb{R}^d$, we introduce the notation $x=(x_1,x')$, where x_1 is the first coordinate of x and x' the last $(d-1)$th ones. By symmetry we consider only the side $x_1=0$. Denote by b_1^ε the function satisfying the same equation as b^ε on $\Omega\times V$, with the same condition on the boundary where $x_1=0$ and vanishing boundary conditions on the other sides of Ω. For conciseness, we do not write the equation satisfied by b_1^ε. Since $w_0=0$ on $\partial\Omega$, we obtain for $x=(0,x')$ the relation

$$\zeta_3^\varepsilon(x,\mu)=\kappa_1\left(\frac{x}{\varepsilon},\mu\right)\frac{\partial w_0}{\partial x_1}(x)-w_{10}(x).$$

We are interested in the asymptotic behavior of b_1^ε. Let ε go to 0 and consider a point $(0,x')$ on the side $x_1=0$. We perform a stretching around the point $(0,x')$ in

order to obtain an equation for $b_1^\varepsilon(\varepsilon x,\mu)$. In the limit $\varepsilon=0$ we obtain formally the following equation for $b_\infty(y,\mu)$, where x' is a parameter:

$$\begin{aligned} -\mu\cdot\nabla_y\frac{(\psi_\infty^+)^2}{\Sigma}\mu\cdot\nabla_y b_\infty + Qb_\infty &= 0 \qquad \text{in } \mathbb{R}^+\times\mathbb{R}^{d-1}\times V,\\ b_\infty - \frac{\psi_\infty^+}{\psi_\infty\Sigma}\mu\cdot\nabla_y b_\infty &= \kappa_1(y,\mu)\left[\frac{\partial w_0}{\partial x_1}(0,x')\right] - [w_{10}(0,x')] \qquad \text{on } \Gamma_-. \end{aligned} \tag{57}$$

This problem is similar to the Milne problem presented, for instance, in [14, 20, 33]. For some particular matching conditions between w_{10} and $\frac{\partial w_0}{\partial x_i}$ we obtain that b_∞ decays exponentially fast as $x_1\to+\infty$. This problem is studied in Lemma 6.2, given in section 6. We deduce from this lemma that (57) admits a unique solution in $W^2(\mathbb{R}^+\times\mathbb{R}^{d-1}\times V)$, which converges, as x_1 goes to infinity, to a constant term $G(\kappa_1\frac{\partial w_0}{\partial x_1}-w_{10}) = G(\kappa_1)\frac{\partial w_0}{\partial x_1} - w_{10}$, where $G(\kappa_1)=L$ is a constant. We obtain that the layer b_∞ vanishes when $x_1\to\infty$ provided $G(\kappa_1\frac{\partial w_0}{\partial x_1}-w_{10})=0$. In other words,

$$L\frac{\partial w_0}{\partial n} + w_{10} = 0. \tag{58}$$

This corresponds to the boundary conditions (32).

Let us now prove that b^ε is of order $O(\varepsilon^{1/2})$ in $L^2(\Omega\times V)$ when (58) is satisfied. We define

$$\gamma^0 = \{x\in\partial\Omega \text{ s.t. } x_1=0\}.$$

The variational formulation for b_1^ε is given for every test function $v\in\mathcal{V}$ by

$$\begin{aligned} &\int_\Omega\int_V \frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}\mu\cdot\nabla b_1^\varepsilon\,\mu\cdot\nabla v\,d\mu dx + \frac{1}{\varepsilon^2}\int_\Omega\int_V (Q_\varepsilon b_1^\varepsilon)v\,d\mu dx\\ &+\frac{1}{\varepsilon}\int_{\partial\Omega}\int_V |\mu\cdot n|h_\varepsilon b_1^\varepsilon v\,d\mu d\sigma = \frac{1}{\varepsilon}\int_{\gamma^0}\int_V |\mu\cdot n|h_\varepsilon\left(\kappa_1\left(\frac{x}{\varepsilon},\mu\right)-L\right)\partial_1 w_0(x')v\,d\mu d\sigma. \end{aligned} \tag{59}$$

Let us introduce

$$d^\varepsilon(x_1,x',\mu) = -\partial_1 w_0(x')U(x_1)b\left(\frac{x}{\varepsilon},\mu\right), \tag{60}$$

where $U(x_1)\in C^\infty(\mathbb{R}^+)$ is such that $U(0)=1$ and $U(x_1)=0$ for $x_1\geq 1$, and where b is the solution to (66) with $g=\kappa_1-L$, $S=0$, and $T=0$. According to Lemma 6.2, b is exponentially decaying to 0. We have that

$$\begin{aligned} A_\varepsilon d^\varepsilon &= \tilde S - \mu\cdot\nabla\tilde T \quad \text{in } X,\\ d^\varepsilon - \frac{\varepsilon\psi_\varepsilon^+}{\psi_\varepsilon\Sigma^\varepsilon}\mu\cdot\nabla d^\varepsilon &= \left(L-\kappa_1\left(\frac{x}{\varepsilon},\mu\right)\right)\partial_1 w_0(x') \quad \text{on } \Gamma_-^0, \end{aligned} \tag{61}$$

with the notation of section 6. Here we have

$$\begin{aligned} \tilde T &= 2b\left(\frac{x}{\varepsilon},\mu\right)\mu\cdot\nabla(-\partial_1 w_0(x')U(x_1))\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon},\\ \tilde S &= (\mu\cdot\nabla)^2(-\partial_1 w_0(x')U(x_1))\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon}b\left(\frac{x}{\varepsilon},\mu\right). \end{aligned}$$

According to the Y'-periodicity of b (see section 6 for the notation), we obtain the following variational formulation for d^ε:

$$\begin{aligned}
&\int_X \frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon} \mu\cdot\nabla d^\varepsilon\, \mu\cdot\nabla v\, d\mu dx + \frac{1}{\varepsilon^2}\int_X (Q_\varepsilon d^\varepsilon) v\, d\mu dx + \frac{1}{\varepsilon}\int_{\Gamma^0} |\mu\cdot n| h_\varepsilon d^\varepsilon v\, d\mu d\sigma \\
&= \frac{1}{\varepsilon}\int_{\Gamma^0} \left[|\mu\cdot n| h_\varepsilon \left(\kappa_1\left(\frac{x}{\varepsilon},\mu\right) - L\right) \partial_1 w_0(x') v - \varepsilon(\mu\cdot n)\tilde{T} v \right] d\mu d\sigma \\
&\quad + \int_X (\tilde{S} v + \tilde{T}\mu\cdot\nabla v) d\mu dx. \qquad (62)
\end{aligned}$$

Let us define $\delta^\varepsilon = d^\varepsilon - b_1^\varepsilon$. Since $w_0 = 0$ on $\partial\Omega$, we have $d^\varepsilon = 0$ on $\partial\Omega\backslash\gamma^0$. Therefore, subtracting (59) from (62), we obtain that

$$\begin{aligned}
&\int_\Omega \int_V \Big(\frac{(\psi_\varepsilon^+)^2}{\Sigma^\varepsilon} \mu\cdot\nabla\delta^\varepsilon\, \mu\cdot\nabla v\, d\mu dx + (Q_\varepsilon \delta^\varepsilon) v \Big) d\mu dx + \frac{1}{\varepsilon}\int_{\partial\Omega}\int_V |\mu\cdot n| h_\varepsilon \delta^\varepsilon v\, d\mu d\sigma \\
&= -\int_{\gamma^0}\int_V (\mu\cdot n)\tilde{T} v d\mu d\sigma + \int_\Omega \int_V (\tilde{S} v + \tilde{T}\mu\cdot\nabla v) d\mu dx. \qquad (63)
\end{aligned}$$

Because b decays exponentially fast, we deduce that

$$\left\| b\left(\frac{x}{\varepsilon},\mu\right)\right\| \le C\sqrt{\varepsilon}.$$

Therefore, the source terms $\tilde{T}$ and $\tilde{S}$ satisfy the same bound. Choose $v = \delta^\varepsilon$ in this expression. We have

$$\|\mu\cdot\nabla\delta^\varepsilon\|^2 + \frac{1}{\varepsilon}\|\delta^\varepsilon\|^2_{L^2(\Gamma_-,d\xi)} \le C\|\delta^\varepsilon\|_{L^2(\Gamma_-,d\xi)} + C\sqrt{\varepsilon}(\|\delta_\varepsilon\| + \|\mu\cdot\nabla\delta^\varepsilon\|).$$

Recalling the Poincaré inequality (42), we easily obtain that

$$\|\delta^\varepsilon\| + \|\mu\cdot\nabla\delta^\varepsilon\| + \varepsilon^{-1/2}\|\delta^\varepsilon\|_{L^2(\Gamma_-,d\xi)} \le C\sqrt{\varepsilon}.$$

Since $b_1^\varepsilon = d^\varepsilon - \delta^\varepsilon$, we deduce that $\|b_1^\varepsilon\| \le C\sqrt{\varepsilon}$. This concludes the proof of the theorem. ∎

6. The conservative multidimensional Milne problem in a periodic half space. In this section, we state the lemma we used in the proof of Theorem 3.11. The study of problem (57) plays a crucial role in the construction of the first-order corrector. For homogeneous problems in slab geometry, it is often referred to as the conservative Milne problem. It has been studied by many authors in the setting of the first-order integrodifferential form of the transport equation. First introduced by astrophysicists [18], it has been analyzed in [14] using probabilistic techniques. More recently, it has been revisited in [11] using results of functional analysis. The tools introduced by these authors, such as the exponential decay analysis, are used here. The specific structure of the second-order formulation in a genuine multidimensional geometry necessitates employing additional methods.

Although we do not need it in the proof of convergence, we also study the half space problem with an exponentially decaying source term away from the boundary. Similar results were derived in diffusion theory for first-order correctors of source problems [13].

Let us introduce some notation. For $x \in \mathbb{R}^d$, we have $x = (x_1, x')$. We denote by $Y' = (0,1)^{d-1}$ and for $A \in \mathbb{R}^+$, $\Gamma^A = \{A\} \times Y' \times V$. Let $X_{AB} = (A, B) \times Y' \times V$ and $X = X_{0,+\infty}$. Its boundary is decomposed as $\partial X = \Gamma^\# \cup \Gamma^0$. The functional space $D(X)$ is defined by

$$\begin{aligned} &D(X) = \{u \in W^2_{\#,loc}(X), \|u\|_{D(X)} < +\infty\}, \\ &\|u\|_{D(X)} = \|\mu \cdot \nabla u\|_{L^2(X)} + \|u - \langle u\rangle\|_{L^2(X)} + \|u\|_{L^2(\Gamma^0_-, d\xi)}. \end{aligned} \tag{64}$$

Here $W^2_{\#,loc}(X)$ is the space of functions of $W^2_{loc}(X)$ that are Y'-periodic with respect to x' and Γ^0_- is the set of incoming boundary conditions for X at side Γ^0.

LEMMA 6.1. *The space $D(X)$ defined in* (64) *is Hilbert.*

Proof. We have to prove that the seminorm given in (64) is a norm for $D(X)$. Actually we prove that

$$\|u\|_{X_{0M}} \leq CM(\|\mu \cdot \nabla u\|_{L^2(X)} + \|u - \langle u\rangle\|_{L^2(X)} + \|u\|_{L^2(\Gamma^0_-, d\xi)}).$$

This is enough to obtain the local integrability. We first replace the domain of integration of the velocity by a smaller set of positive directions. Let B be a ball centered at μ such that $\mu_1 = 0$ of a given radius. Let $\theta \in V = S^{d-1}$ and B_θ be the image of B by a rotation R_θ of angle θ. Let $f \in L^2(S^{d-1})$. We have

$$\begin{aligned} &\int_{B_\theta} f^2(\mu) d\mu = \int_{B_\theta} [(f - \langle f\rangle) - (R_{-\theta} f - \langle f\rangle) + R_{-\theta} f]^2 d\mu \\ \leq\ & 3\int_B f^2(\mu) d\mu + 6\int_V (f - \langle f\rangle)^2 d\mu. \end{aligned}$$

Let $B = \{\mu \in V \text{ s.t. } \mu_1 > 1/2\}$, for instance. A finite number of angles θ is sufficient so that the union of the corresponding B_θ covers V. Then we have that

$$\|u\|_{L^2(X_{0M})} \leq C(\|u\|_{(0,M)\times Y' \times B} + \|u - \langle u\rangle\|_{L^2(X)}),$$

where C is a universal constant. Since u is bounded at the boundary Γ^0_- and $\mu \cdot \nabla u$ is bounded in $L^2(X)$, it is not difficult to obtain a bound for $\|u\|_{(0,M)\times Y' \times B}$. We have that

$$u(x,\mu) = \int_0^{d(x,\mu)} \mu \cdot \nabla u(x - s\mu, \mu) ds + u(\overline{x}, \mu), \quad (x,\mu) \in (0,M) \times Y' \times B.$$

Here $d(x,\mu)$ is the distance between $x \in \mathbb{R}^+ \times \mathbb{R}^{d-1}$ and the surface $x_1 = 0$ in the direction $-\mu$ and $\overline{x}$ the point of the interface defined by $\overline{x} = x - d(x,\mu)\mu$. We deduce from the Cauchy–Schwartz inequality that

$$|u(x,\mu)|^2 \leq CM \int_0^{d(x,\mu)} (\mu \cdot \nabla u)^2 (x - s\mu, \mu) ds + 2|u(\overline{x}, \mu)|^2, \tag{65}$$

where C is a universal constant. For $\mu \in B$ given, we integrate and obtain

$$\int_0^M \int_{Y'} |u(x,\mu)|^2 dx \leq CM^2 \|\mu \cdot \nabla u\|^2_{L^2(X)} + CM\|u\|^2_{L^2(\Gamma^0_-, d\xi)}.$$

It remains to integrate this relation over B to complete the proof of the lemma. ∎

We can now state the main result of this section.

LEMMA 6.2. *Let $g \in L^2(\Gamma^0_-, d\xi)$, $S \in L^2(X)$, $T \in L^2(X)$ be sufficiently regular so that its trace on γ^0 is defined, and $\nu > 0$. We assume that $\check{S} = S$ and $\check{T} = -T$. We denote by $\mathcal{N} = \|g\|_{L^2(\Gamma^0_-,d\xi)} + \|S\|_{L^2(X)} + \|T\|_{L^2(X)} + \|T\|_{L^2(\gamma^0_-,d\xi)}$. There exists a unique solution in $D(X)$ to the following boundary layer problem:*

$$\begin{aligned} -\mu\cdot\nabla\frac{(\psi^+_\infty)^2}{\Sigma}\mu\cdot\nabla b_\infty + Qb_\infty &= Se^{-\nu x_1} - \mu\cdot\nabla(Te^{-\nu x_1}) \quad \text{in } X, \\ b_\infty - \frac{\psi^+_\infty}{\psi_\infty\Sigma}\mu\cdot\nabla b_\infty &= g \quad \text{on } \Gamma^0_-. \end{aligned} \tag{66}$$

The solution b_∞ satisfies $\check{b}_\infty = b_\infty$. Moreover, there exist three linear forms G and H and J such that b_∞ decays exponentially as $x_1 \to \infty$ to a constant $L = G(g) + H(S) + J(T)$ satisfying $|L| \le C\mathcal{N}$. More specifically, we have

$$\begin{aligned} &\|b_\infty - L\|_{L^2(X_{M,\infty})} + \|\mu\cdot\nabla b_\infty\|_{L^2(X_{M,\infty})} \\ &+ \|(\mu\cdot\nabla)^2 b_\infty\|_{L^2(X_{M,\infty})} + \|Qb_\infty\|_{L^2(X_{M,\infty})} \le C\mathcal{N}\, e^{-\beta M/2}, \end{aligned} \tag{67}$$

where C and β are constants independent of b_∞ and M. Notice that $G(1) = 1$.

Proof. (i) *Existence and uniqueness of the solution.* Let $v \in D(X)$ be a test function with compact support. Multiplying (66) by v and integrating by parts yields

$$\begin{aligned} &\int_X \left(\frac{(\psi^+_\infty)^2}{\Sigma}\mu\cdot\nabla b_\infty \mu\cdot\nabla v + Qb_\infty v\right) + \int_{\Gamma^0}(\psi^+_\infty\psi_\infty)|\mu\cdot n|(b_\infty - g)v \\ = &\int_X e^{-\nu x_1}Sv + e^{-\nu x_1}T\mu\cdot\nabla v - \int_{\Gamma^0}(\mu\cdot n)Tv. \end{aligned} \tag{68}$$

The measures of integration are $d\mu dx$ on X_{AB} and $d\mu dx'$ on Γ^A. We drop them for simplicity. By density, the same equation holds true for every $v \in D(X)$. With obvious notation, problem (66) is equivalent to

$$a(b_\infty, v) = \mathcal{F}(v) \quad \forall v \in D(X). \tag{69}$$

The form a is clearly bicontinuous and coercive in $D(X)$. We deduce from the proof of Lemma 6.1 that

$$\left|\int_X e^{-\nu x_1}T\mu\cdot\nabla v\right| + \left|\int_X e^{-\nu x_1}Sv\right| + \left|\int_{\Gamma^0}(\psi^+_\infty\psi_\infty)|\mu\cdot n|g\,v\right| + \left|\int_{\Gamma^0}|\mu\cdot n|T\,v\right| \le C\mathcal{N}\|v\|_{D(X)},$$

where C is a constant independent of g, S, and $v \in D(X)$. Therefore, the form $\mathcal{F}$ is linear in $D(X)$. By the Lax–Milgram lemma, there exists a unique solution to (66) in $D(X)$. From (66), we also deduce that $\mu\cdot\nabla_y b_\infty \in D(X)$.

(ii) *Convergence of b_∞ when $x_1 \to \infty$.* Let be $A < B$. Integrating (66) by parts yields

$$\begin{aligned} &\int_{X_{AB}}\left(\frac{(\psi^+_\infty)^2}{\Sigma}\mu\cdot\nabla b_\infty\mu\cdot\nabla v + Qb_\infty v\right) \\ &- \int_{\Gamma^A\cup\Gamma^B}\left(\frac{(\psi^+_\infty)^2}{\Sigma}\mu\cdot\nabla b_\infty - Te^{-\nu x_1}\right)(\mu\cdot n)v = \int_{X_{AB}} e^{-\nu x_1}Sv + e^{-\nu x_1}T\mu\cdot\nabla v. \end{aligned} \tag{70}$$

Let us choose $v = 1$ in this variational formulation. Recalling that $\mu \cdot n = \mu_1$ on Γ^B and $\mu \cdot n = -\mu_1$ on Γ^A, we have for $B = \infty$,

$$\int_{\Gamma^A} \frac{(\psi_\infty^+)^2}{\Sigma} \mu \cdot \nabla b_\infty \mu_1 = \int_{X_{A,\infty}} e^{-\nu x_1} S - \int_{\Gamma^A} (\mu \cdot n) e^{-\nu A} T. \tag{71}$$

Choosing $v = b_\infty$ now yields

$$\begin{aligned}
&\int_{\Gamma^A} \frac{(\psi_\infty^+)^2}{\Sigma} \mu \cdot \nabla b_\infty \mu_1 b_\infty \\
&= \int_{X_{A,\infty}} \left(-\frac{(\psi_\infty^+)^2}{\Sigma} (\mu \cdot \nabla b_\infty)^2 - (Q b_\infty) b_\infty + e^{-x_1 \nu} S b_\infty + e^{-x_1 \nu} T \mu \cdot \nabla b_\infty \right) \\
&- \int_{\Gamma^A} (\mu \cdot n) e^{-\nu A} T b_\infty \le \int_{X_{A,\infty}} (e^{-x_1 \nu} S b_\infty + e^{-x_1 \nu} T \mu \cdot \nabla b_\infty) - \int_{\Gamma^A} (\mu \cdot n) e^{-\nu A} T b_\infty.
\end{aligned}$$

We deduce from this equality, Lemma 4.1, and a classical trace theorem [16, 17] that

$$e^{A\nu/2} \int_{\Gamma^A} \frac{(\psi_\infty^+)^2}{\Sigma} \mu \cdot \nabla b_\infty \mu_1 b_\infty \le \|b_\infty\|_{D(X)} \mathcal{N}. \tag{72}$$

Let us now choose $v = b_\infty e^{\beta x_1}$ on X_{0B}, where $0 < \beta < \nu/2$ will be determined later. We obtain that

$$\begin{aligned}
&\int_{X_{0B}} \left(\frac{(\psi_\infty^+)^2}{\Sigma} (\mu \cdot \nabla b_\infty)^2 e^{\beta x_1} + \frac{(\psi_\infty^+)^2}{\Sigma} \mu \cdot \nabla b_\infty \mu_1 b_\infty \beta e^{\beta x_1} + (Q b_\infty) b_\infty e^{\beta x_1} \right) \\
&+ \int_{\Gamma^0} (\psi_\infty^+ \psi_\infty) |\mu \cdot n| b_\infty^2 \\
&= \int_{\Gamma^0} (\psi_\infty^+ \psi_\infty) |\mu \cdot n| b_\infty g + \int_{\Gamma^B} \frac{(\psi_\infty^+)^2}{\Sigma} \mu \cdot \nabla b_\infty \mu_1 b_\infty e^{\beta B} + \int_{X_{0B}} e^{-(\nu-\beta)x_1} S b_\infty \\
&+ \int_{X_{0B}} e^{-(\nu-\beta)x_1} T (\mu \cdot \nabla b_\infty + \mu_1 \beta b_\infty) - \int_{\Gamma^0} T (\mu \cdot n) b_\infty - \int_{\Gamma^B} e^{-(\nu-\beta)B} T (\mu \cdot n) b_\infty.
\end{aligned}$$

We deduce from (72) and Lemma 4.1 that there exist positive constants η and C independent of B such that

$$\int_{X_{0B}} \left[\frac{(\psi_\infty^+)^2}{\Sigma} (\mu \cdot \nabla b_\infty)^2 + \beta \frac{(\psi_\infty^+)^2}{\Sigma} \mu \cdot \nabla b_\infty \mu_1 b_\infty + \eta |b_\infty - \langle b_\infty \rangle|^2 \right] e^{\beta x_1} \le C \mathcal{N}^2. \tag{73}$$

The second term in the left-hand side in (73) has to be estimated in terms of the two other ones, which are positive. Let us introduce the following average operators, defined for every regular function u by

$$m_{\Gamma^B}(u) = \int_{\Gamma^B} u \, d\mu dx', \quad m_M(u) = \int_{X_{M,M+1}} u \, d\mu dx, \quad m_{Y_M}(u) = \int_{Y_M} u \, dx, \tag{74}$$

where $Y_M = (M, M+1) \times Y'$ is the Mth cell. Remark that m_{Y_M} is only a spatial averaging operator and that $m_M(u) = \int_V m_{Y_M}(u)$. We now prove that

$$\|u - m_M(u)\|_{L^2(X_{M,M+1})} \le C \left(\|\mu \cdot \nabla u\|_{L^2(X_{M,M+1})} + \|u - \langle u \rangle\|_{L^2(X_{M,M+1})} \right). \tag{75}$$

Let $\mu \in V$ be given. Since the cell Y_M is bounded, we easily check by integration of one-dimensional problems along the direction μ and along its transverse directions that

$$\|u(\cdot,\mu) - (m_{Y_M}(u))(\mu)\|^2_{L^2(Y_M)} \le C\|\mu\cdot\nabla u\|^2_{L^2(Y_M)},$$

where C is a constant independent of μ. Integrating over V yields

$$\|u - m_Y(u)\|_{L^2(X_{M,M+1})} \le C\|\mu\cdot\nabla u\|_{L^2(X_{M,M+1})}.$$

On the other hand we check that

$$\begin{aligned}\|m_{Y_M}(u) - m_M(u)\|^2_{L^2(X_{M,M+1})} &= \int_{X_{M,M+1}}\left(\int_{Y_M}(u- < u >)\right)^2 \\ &\le C\|u - \langle u\rangle\|^2_{L^2(X_{M,M+1})}.\end{aligned}$$

This proves (75). In addition, we obtain that

$$\|u - m_{\Gamma^{x_1}}(u)\|_{L^2(X_{M,M+1})} \le \|u - m_M(u)\|_{L^2(X_{M,M+1})}.$$

Therefore, we have

$$\|u - m_{\Gamma^{x_1}}(u)\|_{L^2(X_{M,M+1})} \le C\left(\|\mu\cdot\nabla u\|_{L^2(X_{M,M+1})} + \|u - \langle u\rangle\|_{L^2(X_{M,M+1})}\right).$$

This inequality holds for all $M \in \mathbb{R}^+$ and since $e^{\beta M} \le e^{\beta x_1} \le e^{\beta}e^{\beta M}$ for $x_1 \in (M, M+1)$, we obtain

$$\|(u - m_{\Gamma^{x_1}}(u))e^{\frac{\beta}{2}x_1}\|_{L^2(X)} \le C(\|(\mu\cdot\nabla u)e^{\frac{\beta}{2}x_1}\|_{L^2(X)} + \|(u - \langle u\rangle)e^{\frac{\beta}{2}x_1}\|_{L^2(X)}). \tag{76}$$

We deduce from (71) that

$$\begin{aligned}&\left|\int_X \frac{(\psi^+_\infty)^2}{\Sigma}\mu\cdot\nabla b_\infty \mu_1 m_{\Gamma^{x_1}}(b_\infty)e^{\beta x_1}\right| \\ =\ &\left|\int_X m_{\Gamma^{x_1}}(b_\infty)e^{\beta x_1}\left(\int_{X_{x_1\infty}} e^{-\nu y_1}S - \int_{\Gamma^{x_1}}(\mu\cdot n)e^{-\nu x_1}T\right)\right| \\ \le\ & C\int_X e^{-(\nu-\beta)x_1}|m_{\Gamma^{x_1}}(b_\infty)|(\|S\|_{L^2(X)} + \|T\|_{L^2(X)}) \le C\mathcal{N}^2.\end{aligned} \tag{77}$$

Here and in what follows, C denotes a constant independent of b_∞. Then according to (76) we can rewrite (73) and obtain for some constants $\eta > 0$ and $\theta > 0$,

$$\begin{aligned}&\int_X (\mu\cdot\nabla b_\infty)^2 e^{\beta x_1} - \theta\beta\int_X |(\mu\cdot\nabla b_\infty)\,(b_\infty - m_{\Gamma^{x_1}}(b_\infty))|\, e^{\beta x_1} \\ &+\eta\int_X (b_\infty - m_{\Gamma^{x_1}}(b_\infty))^2 e^{\beta x_1} \le C\mathcal{N}^2.\end{aligned} \tag{78}$$

Choosing now $\beta > 0$ small enough, we deduce from the Cauchy–Schwartz inequality that

$$\int_X (\mu\cdot\nabla b_\infty)^2 e^{\beta x_1} + \int_X (b_\infty - m_{\Gamma^{x_1}}(b_\infty))^2 e^{\beta x_1} \le C\mathcal{N}^2. \tag{79}$$

Owing to (73) and (77), we deduce from (79) and the Cauchy–Schwartz inequality that

$$\int_X |b_\infty - \langle b_\infty \rangle|^2 \, e^{\beta x_1} \leq C\mathcal{N}^2.$$

Therefore, we obtain

$$\int_X (\mu \cdot \nabla b_\infty)^2 e^{\beta x_1} + \int_X |b_\infty - \langle b_\infty \rangle|^2 \, e^{\beta x_1} \leq C\mathcal{N}^2. \tag{80}$$

It follows from (75) and (80) that

$$\|u - m_M(u)\|_{L^2(X_{M,M+1})} \leq Ce^{-\frac{\beta M}{2}}\mathcal{N},$$

and averaging over $x_1 \in (M, M+2)$ we have

$$\begin{aligned}\left\| u - \frac{m_M(u) + m_{M+1}(u)}{2} \right\|_{L^2(X_{M,M+1})} &\leq \left\| u - \tfrac{m_M(u)+m_{M+1}(u)}{2} \right\|_{L^2(X_{M,M+2})} \\ &\leq Ce^{-\frac{\beta M}{2}}\mathcal{N}.\end{aligned}$$

Then we readily show that

$$|m_{M+1}(u) - m_M(u)| \leq Ce^{-\frac{\beta M}{2}}\mathcal{N}.$$

This proves that $M \mapsto m_M(u)$ converges exponentially fast to a constant L by linearity of the transport equation. Moreover, due to the exponential rate of convergence, we have that

$$|L| \leq |G(g)| + |H(S)| + |J(T)| \leq C\mathcal{N}.$$

Let us introduce

$$m(u) = m_M(u) \qquad \text{for } x_1 \in (M, M+1).$$

Then $m(b_\infty)$ is an element of $L^2(X)$ that converges to L and

$$\|[b_\infty - m(b_\infty)]e^{\frac{\beta x_1}{2}}\|_{L^2(X)} \leq C\mathcal{N}.$$

Hence we can conclude that b_∞ converges to L in the following sense:

$$\|b_\infty - L\|_{L^2(X_{M,\infty})} \leq Ce^{-\frac{\beta M}{2}}\mathcal{N}.$$

The same results are derived for $\mu \cdot \nabla_y b$ and for Qb. This completes the proof of the lemma. ∎

7. Numerical application and conclusion. It goes beyond the scope of this paper to present an extensive numerical application of the theory we have described. We refer to [8, 9] for a more detailed analysis. However, we show the convergence of the eigenvalues for the homogenization of a core composed of $N \times N$ uranium assemblies, which are common in thermal reactors. In its two-dimensional approximation, each assembly is made of 17×17 fuel pins or control rods. We do not describe them here and refer to [9] for the details. We present the convergence of the eigenvalue

TABLE 1
Reference and reconstructed k_{eff} for cores composed of $N \times N$ uranium assemblies.

N	Reference	k_{eff}, $L = 0$	Error (10^{-5})	k_{eff}, $L = 0.71$	Error (10^{-5})
5	0.46598	0.45542	1056	0.46543	55
10	0.54521	0.54334	187	0.54515	6
20	0.57114	0.57089	25	0.57114	0
40	0.57825	0.57821	4	0.57825	0
∞	0.58070				

$k_{eff} = 1/\lambda_1^\varepsilon$ for different values of the number of assemblies $N = 1/\varepsilon$. In Table 1 are gathered the exact multiplication factor k_{eff}, the associated diffusion approximation given by $(\lambda_\infty + \varepsilon^2\nu_1)^{-1}$ corresponding to an extrapolation length $L = 0$, and the diffusion approximation $(\lambda_\infty + \varepsilon^2\omega_1^\varepsilon)^{-1}$ that corresponds to an extrapolation length of approximately $L = 0.71$. The exact extrapolation length has not been computed. Instead, we took the value of the extrapolation length in homogeneous media. This can be physically justified by the fact that uranium assemblies are not very heterogeneous.

As expected from the theory, the convergence of the eigenvalues without accounting for the leakage is of order ε^3. A numerical estimate is given by

$$v \;=\; \frac{\ln \frac{e_{20}}{e_{10}}}{\ln \frac{20}{10}} \simeq 2.9,$$

which is in good agreement with the theoretical value. When we account for the neutron leakage, the rate of convergence is too fast to be estimated numerically. Indeed two-dimensional computations are already highly demanding, especially for a number of assemblies equal to 20×20 and 40×40, which correspond to 115,600 and 462,400 rods, respectively. Hence we obtained an accuracy of 10^{-5} for the eigenvalues. Nevertheless the gain obtained by adding an extrapolation length is clear according to the numerical results presented in Table 1. We do not give here the reconstructed fluxes for both methods. However, in both cases the shape of the exact fluxes is well reproduced by the reconstructed fluxes [9]. In the case of a zero extrapolation length, the reconstructed flux is shifted downward. This is explained by the Dirichlet boundary conditions, which do not account for the neutron leakage. The error made is of the order of 5% for a core composed of 10×10 assemblies (usual cores have approximately 150 assemblies). When Robin-like boundary conditions are imposed for the diffusion approximation, the reconstructed flux has the correct shape, up to an error of less than 1%. Fine resolution close to the boundary also allows us to see the exponential decay of the boundary layer [9].

In this paper, we have presented the homogenization of the transport criticality eigenvalue problem and a numerical experiment for a one-velocity periodic neutron transport problem. The generalization to multigroup and anisotropic equations is straightforward and does not require new techniques [4, 5]. The results can also be readily extended [8] to cores that are periodic up to a perturbation of order ε^2, i.e., for instance, $\Sigma(\frac{x}{\varepsilon}, \mu) + \varepsilon^2\Sigma'(x, \frac{x}{\varepsilon}, \mu)$. However, the extension to genuinely nonperiodic cores is still open. Periodicity is needed in order to obtain separation of slow and fast variables. It would be interesting, theoretically as well as practically, to understand the homogenization of locally periodic domains. The analysis of boundary layers is replaced by that of interface layers, for which no theory is available at present.

Appendix A. Equivalence between the first-order and even parity flux formulations. A linear integrodifferential Boltzmann equation is usually solved to

model neutron transport. In the simplified setting of one-velocity and isotropic source problems, it is given as follows. Given the cross sections Σ and Σ_s and the source term q, we want to find the neutron angular flux $\psi(x,\mu)$ satisfying

$$\begin{aligned} \mu\cdot\nabla\psi(x,\mu)+\Sigma(x)\psi(x,\mu) &= \Sigma_s(x)\int_V \psi(x,\mu')d\mu'+q(x) \quad \text{in } \Omega\times V,\\ \psi(x,\mu) &= 0 \quad \text{on } \Gamma_-. \end{aligned} \tag{81}$$

The assumption of one-velocity neutrons is not essential in order to derive the even parity flux formulation, and the generalization to multigroup problems is straightforward. However, isotropy of the scattering operator can be hardly avoided, even if complicated generalizations exist for nonisotropic media. This restricts the use of the even parity formulation, although isotropy is a correct approximation in nuclear reactor computations.

The even parity flux formulation consists in deriving a second-order integrodifferential equation for the *even parity flux*, which is defined by

$$\psi^+(x,\mu)=\frac{\psi(x,\mu)+\psi(x,-\mu)}{2}. \tag{82}$$

The main features of this formulation are that the second-order differential operator allows for the use of a variational formulation, which is convenient in theory as well as for numerical implementations. For instance, standard finite element methods can be used as in the resolution of elliptic problems [28, Chapter 6]. Let us also emphasize that the number of angular directions, and therefore the computational cost, has been divided by two.

The derivation of the even parity flux equations from the first-order integrodifferential equation is done as follow. Let us first introduce the *odd parity flux*

$$\psi^-(x,\mu) = \frac{\psi(x,\mu)-\psi(x,-\mu)}{2}$$

and the *scalar flux*

$$\phi(x) = \int_V \psi(x,\mu')d\mu' = \int_V \psi^+(x,\mu')d\mu'.$$

Then (81) written at points (x,μ) and $(x,-\mu)$ yields

$$(\mu\cdot\nabla+\Sigma)(\psi^++\psi^-)=\Sigma_s\phi+q, \quad (-\mu\cdot\nabla+\Sigma)(\psi^+-\psi^-)=\Sigma_s\phi+q.$$

Summing and subtracting these equalities, we obtain that

$$\mu\cdot\nabla\psi^-+\Sigma\psi^+ = \Sigma_s\phi+q, \qquad \mu\cdot\nabla\psi^++\Sigma\psi^- = 0. \tag{83}$$

Since the total cross section Σ is positive, we have

$$\psi^-=-\frac{1}{\Sigma}\mu\cdot\nabla\psi^+. \tag{84}$$

It remains to insert this expression in (83) and obtain

$$-\mu\cdot\nabla\frac{1}{\Sigma(x)}\mu\cdot\nabla\psi^+(x,\mu)+\Sigma(x)\psi^+(x,\mu)=\Sigma_s(x)\int_V\psi^+(x,\mu')d\mu'+q(x). \tag{85}$$

From the boundary conditions for ψ, we deduce that

$$\begin{array}{rcll} \psi^+ + \psi^- & = & 0 & \text{on} \quad \Gamma_- = \{(x,\mu) \in \partial\Omega \times V;\ \mu \cdot n(x) < 0\}, \\ \psi^+ - \psi^- & = & 0 & \text{on} \quad \Gamma_+ = \{(x,\mu) \in \partial\Omega \times V;\ \mu \cdot n(x) > 0\}. \end{array}$$

Then, according to (84), we have

$$\psi^+ - \frac{1}{\Sigma}\mu\cdot\nabla\psi^+ = 0 \text{ on } \Gamma_-, \quad \psi^+ + \frac{1}{\Sigma}\mu\cdot\nabla\psi^+ = 0 \text{ on } \Gamma_+. \tag{86}$$

Remark that due to the symmetry of the even parity flux, these two boundary conditions are redundant, so we need to mention only one of them. Equations (85) and (86) are called the even parity flux formulation of the neutron transport. With the assumption of subcriticality given by

$$\Sigma(x) - \Sigma_s(x) \geq \eta > 0,$$

problem (81) admits a unique solution in $W^2(\Omega \times V)$ (see, e.g., [20, Chapter 21]). Then from the definition (82), we obtain a weak solution of (85)–(86) in $W^2(\Omega \times V)$. The even parity flux ψ^+ also belongs to $\mathcal{V}$, defined by (40). This follows from a trace theorem [16, 17], which ensures that $\psi|_{\Gamma_+} \in L^2(\Gamma_+, d\xi)$ when $\psi \in W^2(\Omega \times V)$ and $\psi = 0$ on Γ_-.

Assume now that $\psi^+ \in \mathcal{V}$ is a weak solution of (85)–(86) and define $\psi = \psi^+ - \frac{1}{\Sigma}\mu\cdot\nabla\psi^+$. From (85), we easily deduce that

$$\mu\cdot\nabla(\psi - \psi^+) + \Sigma\psi^+ = \Sigma_s \int_V \psi + q,$$

and then (81), since (86) clearly implies that $\psi = 0$ on Γ_-. From (81) we deduce that $\psi \in W^2(\Omega \times V)$, and both formulations are equivalent. The same equivalence holds true for eigenvalue problems in bounded or periodic domains.

Appendix B. Some results on operator convergence. The results presented here are derived from [19]. Let X be a Banach space and $\mathcal{L}(X)$ the set of bounded linear operators from X to X. Let $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of operators in $\mathcal{L}(X)$. Then T_n is said to converge *compactly* to T if

- for all $x \in X$, $T_n x \to Tx$ as $n \to \infty$,
- for any bounded sequence $\{x_n\}_{n\in\mathbb{N}}$ with $\|x_n\| \leq 1$, the sequence $\{(T - T_n)x_n\}_{n\in\mathbb{N}}$ is relatively compact in X.

Then we have the following result.

THEOREM B.1. *Let $\{T_n\}_{n\in\mathbb{N}}$ be a sequence of operators in $\mathcal{L}(X)$ converging compactly to T. Let $\sigma(T)$ and $\sigma(T_n)$ be the spectra of T and T_n, respectively. Let λ be an isolated eigenvalue of T of finite multiplicity and let Γ be a closed Jordan curve in the complex plane around λ and isolating λ such that the domain Δ enclosed by Γ contains no other point of the spectrum $\sigma(T)$ than λ. Then $\sigma(T_n) \cap \Delta$ contains a number of eigenvalues equal to the multiplicity of λ provided n is large enough.*

Moreover, let λ_n be a sequence of eigenvalues of T_n converging to λ, and let u_n be a sequence of normalized associated eigenvectors. Then, up to a subsequence, u_n converges to a limit u in X, which is an eigenvector of T associated with λ.

The proof of this theorem relies on Theorem 5.5, p. 232; Proposition 5.6, p. 234; Theorem 5.10, p. 236; Theorem 5.20, p. 244; and Proposition 5.28, p. 24 of [19]. We also have the following error estimate result. Let λ be an eigenvalue of T of finite

multiplicity and Δ defined as above. Let P be the projection on the eigenvectors of T associated with λ and P_n the projection on the eigenvectors associated with eigenvalues of T_n included in Δ. We denote by λ_{in}, $1 \leq i \leq m$, the m eigenvalues counting their multiplicities of T_n in Δ for n large enough. We note $M = PX$ and $M_n = P_n X$. For P and Q two orthogonal projections on X, we define $M = PX$ and $N = QX$ and the distance between M and N by

$$\begin{aligned}\Theta(M,N) &= \max\left(\sup_{x\in M,\,\|x\|=1}\|(1-Q)x\|,\ \sup_{x\in N,\,\|x\|=1}\|(1-P)x\|\right)\\ &= \max\left(\sup_{x\in M,\,\|x\|=1}\operatorname{dist}\,(x,N),\ \sup_{x\in N,\,\|x\|=1}\operatorname{dist}\,(x,M)\right).\end{aligned}$$

Then we prove the following result.

THEOREM B.2. *With the hypotheses and notation of Theorem* B.1*, the following quantities are at least of order* $\|(T-T_n)u\|$*:*

$$\Theta(M,M_n), \qquad \lambda - \frac{1}{m}\sum_{i=1}^{m}\lambda_{in}, \qquad \frac{1}{\lambda} - \frac{1}{m}\sum_{i=1}^{m}\frac{1}{\lambda_{in}} \quad (\text{if } \lambda \neq 0).$$

Moreover if λ *is simple, then* $\lambda - \lambda_n$ *and* $u - u_n$ *are at least of order* $\|(T-T_n)u\|$.

Acknowledgments. The author wishes to thank G. Allaire and G. Papanicolaou for many discussions and comments during the preparation of this paper, and X. Warin for his contribution to the numerical analysis presented here. He also would like to acknowledge helpful comments and remarks from the anonymous referees.

REFERENCES

[1] R. ALEXANDRE AND K. HAMDACHE, *Homogénéisation d'équations cinétiques en milieu perforé*, C. R. Acad. Sci. Paris Sér. I, 313 (1991), pp. 339–344.

[2] G. ALLAIRE, *Homogenization and two-scale convergence*, SIAM J. Math. Anal., 23 (1992), pp. 1482–1518.

[3] G. ALLAIRE AND M. AMAR, *Boundary Layer Tails in Periodic Homogenization*, ESAIM/COCV, 1999.

[4] G. ALLAIRE AND G. BAL, *Homogénéisation d'une équation spectrale de transport neutronique*, C. R. Acad. Sci. Paris Sér. I, 325 (1997), pp. 1043–1048.

[5] G. ALLAIRE AND G. BAL, *Homogenization of the criticality spectral equation in neutron transport*, Math. Model. Anal. Numer., to appear.

[6] G. ALLAIRE AND Y. CAPDEBOSCQ, *Homogenization of a spectral problem in neutronic multigroup diffusion*, Comput. Methods Appl. Mech. Engrg., to appear.

[7] G. ALLAIRE AND F. MALIGE, *Analyse asymptotique spectrale d'un problème de diffusion neutronique*, C. R. Acad. Sci. Paris Sér. I, 324 (1997), pp. 939–944.

[8] G. BAL, *Couplage d'Équations et Homogénéisation en Transport Neutronique*, Ph.D. thesis, de l'Université Paris 6, 1997 (in French).

[9] G. BAL AND X. WARIN, *Numerical Results in the Homogenization of the Criticality Eigenvalue Problem*, preprint, 1998.

[10] C. BARDOS, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre*, Ann. Sci. École Norm. Sup., 4 (1969), pp. 185–233.

[11] C. BARDOS, R. SANTOS, AND R. SENTIS, *Diffusion, approximation and computation of the critical size*, Trans. Amer. Math. Soc., 284 (1984), pp. 617–649.

[12] P. BENOIST, *Théorie du coefficient de diffusion des neutrons dans un réseau comportant des cavités*, Note CEA-R 2278, Commissanat à l'énergie atomique, 1964 (in French).

[13] A. BENSOUSSAN, J. L. LIONS, AND G. PAPANICOLAOU, *Boundary Layer Analysis in Homogeneization of Diffusion Equations with Dirichlet Conditions in the Half Space*, Proc. Internat. Symp. SDE, Kyoto Univ., 1978, pp. 21–40.

[14] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Boundary Layers and Homogenization of Transport Processes*, Res. Inst. Math. Sci., Kyoto Univ., 1979.
[15] J. Bussac and P. Reuss, *Traité de neutronique*, Hermann, Paris, 1978 (in French).
[16] M. Cessenat, *Théorèmes de trace L^p pour des espaces de fonctions de la neutronique*, C. R. Acad. Sci. Paris Sér. I, 299 (1984), pp. 831–834.
[17] M. Cessenat, *Théorèmes de trace pour des espaces de fonctions de la neutronique*, C. R. Acad. Sci. Paris Sér. I, 300 (1985), pp. 89–92.
[18] S. Chandrasekhar, *Radiative Transfer*, Oxford University Press, Cambridge, 1950.
[19] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.
[20] R. Dautray and J. L. Lions, *Analyse Mathématique et Calcul Numérique, Tome* 3, Masson, Paris, 1984.
[21] V. Deniz, *The theory of neutron leakage in reactor lattices*, in Handbook of Nuclear Reactor Calculations, Vol. II, Y. Ronen, ed., 1968, pp. 409–508.
[22] L. Dumas and F. Golse, *Homogenization of transport equations*, SIAM J. Appl. Math., to appear.
[23] E. Frenod and K. Hamdache, *Homogenization of a transport kinetic equation with oscillating potentials*, Proc. Roy. Soc. Edinburgh Set. A, 126 (1996), pp. 1247–1275.
[24] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1977.
[25] F. Golse, P.-L. Lions, B. Perthame, and R. Sentis, *Regularity of the moments of the solution of a transport equation*, J. Funct. Anal., 76 (1988), pp. 110–125.
[26] E. W. Larsen, *Neutron transport and diffusion in inhomogeneous media.* I, J. Math. Phys., 16 (1975), pp. 1421–1427.
[27] E. W. Larsen, *Neutron transport and diffusion in inhomogeneous media.* II, Nuclear Sci. Engrg., 60 (1976), pp. 357–368.
[28] E. E. Lewis and W. F. Miller Jr., *Computational Methods of Neutron Transport*, John Wiley & Sons, New York, 1984.
[29] F. Malige, *Etude Mathématique et Numérique de l'Homogénéisation des Assemblages Combustibles d'un Cœur de Réacteur Nucléaire*, Ph.D. thesis, de l'Ecole polytechnique, 1996 (in French).
[30] J. Pitkäranta, *Estimates for the derivatives of solutions to weakly singular Fredholm integral equations*, SIAM J. Math. Anal., 11 (1980), pp. 952–968.
[31] J. Planchard, *Méthodes mathématiques en neutronique*, Collection de la Direction des Etudes et Recherches d'EDF, Eyrolles, 1995 (in French).
[32] F. Santosa and M. Vogelius, *First-order corrections to the homogenized eigenvalues of a periodic composite medium*, SIAM J. Appl. Math., 53 (1993), pp. 1636–1668.
[33] R. Sentis, *Study of the corrector of the eigenvalue of a transport operator*, SIAM J. Math. Anal., 16 (1985), pp. 151–166.
[34] V. S. Vladimirov, *Mathematical Problems in the One-Velocity Theory of Particle Transport*, Ont. Report AECL 1661, Atomic Energy of Canada Ltd., Chalk River, 1963.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1241–1263

ON THE TWO-DIMENSIONAL GIERER–MEINHARDT SYSTEM WITH STRONG COUPLING*

JUNCHENG WEI† AND MATTHIAS WINTER‡

Abstract. We construct solutions with a single interior condensation point for the two-dimensional Gierer–Meinhardt system with strong coupling. The condensation point is located at a nondegenerate critical point of the diagonal part of the regular part of Green's function for $-\Delta + 1$ under the Neumann boundary condition. Our method is based on the Liapunov–Schmidt reduction for a system of elliptic equations.

Key words. pattern formation, mathematical biology, singular perturbation, strong coupling

AMS subject classifications. Primary, 35B40, 35B45; Secondary, 35J55, 92C15, 92C40

PII. S0036141098347237

1. Introduction.

We study the Gierer–Meinhardt system (see [14]) which models biological pattern formation and can be written as follows (already suitably scaled):

$$\text{(GM)} \qquad \begin{cases} A_t = \epsilon^2 \Delta A - A + \frac{A^p}{H^q}, \quad A > 0 \quad \text{in } \Omega, \\ \tau H_t = D\Delta H - H + \frac{A^r}{H^s}, \quad H > 0 \quad \text{in } \Omega, \\ \frac{\partial A}{\partial \nu} = \frac{\partial H}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \end{cases}$$

Here, the unknowns $A = A(x,t)$ and $H = H(x,t)$ represent the concentrations at a point $x \in \Omega \subset R^N$ and at time t of the biochemicals called activator and inhibitor, respectively; ϵ, τ, D are positive constants; $\Delta := \sum_{j=1}^N \frac{\partial^2}{\partial x_j^2}$ is the Laplace operator in R^N; Ω is a smooth bounded domain in R^N; $\nu(x)$ is the outer normal at $x \in \partial\Omega$. The exponents p, q, r, s are assumed to satisfy the conditions

$$\text{(A)} \qquad 1 < p < \left(\frac{N+2}{N-2}\right)_+, \; q > 0, \; r > 0, \; s \geq 0, \text{ and } 0 < \frac{p-1}{q} < \frac{r}{s+1},$$

where $(\frac{N+2}{N-2})_+ = \frac{N+2}{N-2}$ if $N \geq 3$; $= +\infty$ if $N = 1, 2$. For a related model, see [20].

In numerical simulations of the activator-inhibitor system (GM), it is observed that, when the ratio ϵ^2/D is small, (GM) seems to have stable stationary solutions with the property that the activator concentration is localized around a finite number of points in $\overline{\Omega}$. Moreover, as $\epsilon \to 0$ the pattern exhibits a *"point condensation phenomenon."* By this we mean that the activator concentration is localized in narrower and narrower regions around some points and eventually shrinks to a certain set of points as $\epsilon \to 0$. Hereby the maximum value of the inhibitor concentration diverges to $+\infty$.

*Received by the editors November 17, 1998; accepted for publication March 10, 1999; published electronically October 4, 1999. This work was supported by Stiftung Volkswagenwerk (RiP Program at Oberwolfach) and by Research Grants Council of Hong Kong/Deutscher Akademischer Austauschdienst of Germany (Hong Kong–Germany Joint Research Collaboration).

http://www.siam.org/journals/sima/30-6/34723.html

†Department of Mathematics, The Chinese University of Hong Kong, Shatin, Hong Kong, China (wei@math.cuhk.edu.hk). The research of this author was supported by an Earmarked Grant from Research Grants Council of Hong Kong.

‡Mathematisches Institut A, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany (winter@mathematik.uni-stuttgart.de).

The stationary equation for (GM) is the following system of elliptic equations:

$$\begin{cases} \epsilon^2 \Delta A - A + \frac{A^p}{H^q} = 0, \quad A > 0 \quad \text{in } \Omega, \\ D\Delta H - H + \frac{A^r}{H^s} = 0, \quad H > 0 \quad \text{in } \Omega, \\ \frac{\partial A}{\partial \nu} = \frac{\partial H}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{1.1}$$

Generally speaking, system (1.1) is quite difficult to solve since it has neither a variational structure nor a priori estimates. One way to study (1.1) is to examine the so-called shadow system. Namely, we let $D \to +\infty$ first.

It is known (see [23], [31], [34], [39]) that the study of the shadow system amounts to the study of the following single equation:

$$\begin{cases} \epsilon^2 \Delta u - u + u^p = 0, \quad u > 0 \quad \text{in } \Omega, \\ \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{1.2}$$

Equation (1.2) has a variational structure and has been studied by numerous authors. It is known that (1.2) has both boundary spike solutions and interior spike solutions. For boundary spike solutions, see [5], [9], [10], [15], [17], [22], [25], [29], [30], [31], [39], [44], [46], and the references therein. (When $p = \frac{N+2}{N-2}, N \geq 3$, boundary spike solutions of (1.2) have been studied in [1], [2], [3], [12], [13], [27], etc.) For interior spike solutions, please see [4], [6], [18], [21], [33], [38], [40], [41], [45]. For stability of spike solutions, please see [7], [19], [26], [32], [42], and [43].

In the case when D is finite and not large (this is the so-called strong coupling case), there are only very few results available. For $N = 1$, one can construct spike solutions for all $D \geq 1$; see [37]. In higher dimensions, as far as we know, there are no results yet. (See [8], [28], and [34] for the study of related systems.) In this paper, we consider the case $N = 2$ since it has a particular asymptotic behavior.

Remark. Our approach does not work for dimensions $N \geq 3$ due to a different asymptotic behavior of Green's function of $-\Delta + 1$ with the Neumann boundary condition.

From now on we suppose that $N = 2$. For simplicity, we let $D = 1$.

We construct solutions with a single interior condensation point. It turns out that the condensation points in this case are different from those in the shadow system. We need to introduce some notation. Let $G(P, x)$ be Green's function of $-\Delta + 1$ under the Neumann condition, i.e., G satisfies

$$\begin{cases} -\Delta G + G = \delta_P \quad \text{in } \Omega, \\ \frac{\partial G}{\partial \nu} = 0 \quad \text{on } \partial\Omega, \end{cases}$$

where δ_P is the Dirac delta distribution at point P. It is also known that

$$G(P, x) = K(|x - P|) - H(P, x),$$

where $K(|x|)$ is the fundamental solution of $-\Delta + 1$ in R^2 with singularity at 0 and $H(P, x)$ is C^2 in Ω. It is known that

$$K(r) = -\log r - \mu + O(r) \quad \text{for } r \text{ small}. \tag{1.3}$$

We call $h(P) := H(P, P)$ the diagonal part of $H(P, x)$.

We have Theorem 1.1.

THEOREM 1.1. *Let $P_0 \in \Omega$ be a nondegenerate critical point of $h(P)$. Then for ϵ sufficiently small, problem (1.1) has a solution (A_ϵ, H_ϵ) with the following properties:*

(1) $A_\epsilon(x) = \xi_\epsilon^{q/(p-1)}(w(\frac{x-P_\epsilon}{\epsilon}) + o(1))$ *uniformly for $x \in \bar{\Omega}$, where $\xi_\epsilon > 0$ will be determined later, $P_\epsilon \to P_0$ as $\epsilon \to 0$, and w is the unique solution of the problem*

$$\begin{cases} \Delta w - w + w^p = 0, \quad w > 0 \quad \text{in } R^2, \\ w(0) = \max_{y \in R^2} w(y), \quad w(y) \to 0 \text{ as } |y| \to \infty. \end{cases} \tag{1.4}$$

(2) $H_\epsilon(x) = \xi_\epsilon(1 + O(\frac{1}{|\log \epsilon|}))$ *uniformly for $x \in \bar{\Omega}$.*

(3) $\xi_\epsilon^{s+1-\frac{qr}{p-1}} = (1 + o(1))\epsilon^2 \log \frac{1}{\epsilon} \int_{R^2} w^r$.

Remark. It is known that the solution w to (1.4) is radial, unique, and decays exponentially. (See [16], [24].)

We now outline the proof of Theorem 1.1.

Our method is based on the Liapunov–Schmidt reduction, which was used in [11], [35], and [36] to study semiclassical solutions of the nonlinear Schrödinger equation

$$\frac{h^2}{2}\Delta U - (V - E)U + U^p = 0 \tag{1.5}$$

in R^N, where V is a potential function and E is a real constant. Namely, in [11], [35], and [36] solutions of (1.4) are constructed near a nondegenerate critical point of V provided that h is sufficiently small. Later this method was used in [17], [18], [41], [44], [45], [46] to construct spike solutions for (1.2).

Here we face a system of elliptic equations. Therefore, the process is more complicated. To lay down the basic idea of our proof, we let

$$A_\epsilon = \xi_\epsilon^{q/(p-1)} \bar{A}_\epsilon, \quad H_\epsilon = \xi_\epsilon \bar{H}_\epsilon,$$

where ξ_ϵ is to be chosen later. It is easy to see that system (1.1) is equivalent to the following:

$$\begin{cases} \epsilon^2 \Delta \bar{A}_\epsilon - \bar{A}_\epsilon + \bar{A}_\epsilon^p / \bar{H}_\epsilon^q = 0 \quad \text{in } \Omega, \\ \Delta \bar{H}_\epsilon - \bar{H}_\epsilon + c_\epsilon \bar{A}_\epsilon^r / \bar{H}_\epsilon^s = 0 \quad \text{in } \Omega, \\ \frac{\partial \bar{A}_\epsilon}{\partial \nu} = \frac{\partial \bar{H}_\epsilon}{\partial \nu} = 0 \quad \text{on } \Omega, \end{cases} \tag{1.6}$$

where

$$c_\epsilon = \xi_\epsilon^{\frac{qr}{p-1} - (s+1)}.$$

We fix a point $P \in \Omega$. We rescale

$$\tilde{A}_\epsilon(y) := \bar{A}_\epsilon(P + \epsilon y), \quad x = \epsilon y + P, \quad y \in \Omega_{\epsilon,P} := \{y | P + \epsilon y \in \Omega\}.$$

Then as $\epsilon \to 0$, if we assume that $\bar{H}_\epsilon(P + \epsilon y) \to 1$ in $L^\infty_{loc}(\Omega_{\epsilon,P})$, we have that $\tilde{A}_\epsilon \to V(y)$, where V satisfies

$$\begin{cases} \Delta V - V + V^p = 0, \quad V > 0 \quad \text{in } R^2, \\ V(0) = \max_{y \in R^2} V(y), \quad V \in H^1(R^2). \end{cases}$$

By a uniqueness result it is known that $V(y) = w(y)$, where w is the unique solution of (1.4). (See [16], [24].) Hence

$$\tilde{A}_\epsilon(y) \sim w(y).$$

(Here and thereafter $A \sim B$ means $A = (1 + o(1))B$ as $\epsilon \to 0$ in the corresponding norm.)

To ensure that $\bar{H}_\epsilon(P+\epsilon y) \sim 1$, we note that

$$\bar{H}_\epsilon(P) = \int_\Omega G(P,x)\xi_\epsilon^{\frac{qr}{p-1}-(s+1)} \frac{\bar{A}_\epsilon^r(x)}{\bar{H}_\epsilon^s(x)}dx$$

$$= \epsilon^2 \int_{\Omega_{\epsilon,P}} G(P,P+\epsilon y)\xi_\epsilon^{\frac{qr}{p-1}-(s+1)} \frac{\tilde{A}^r(y)}{\bar{H}_\epsilon^s(P+\epsilon y)}dy$$

(by (1.3), $K(r) = -\log r - \mu + O(r)$ for r small)

$$\sim \xi_\epsilon^{\frac{qr}{p-1}-(s+1)}\epsilon^2 \log\frac{1}{\epsilon}\int_{R^2} w^r(y)\,dy.$$

This suggests that we take

$$\xi_\epsilon^{\frac{qr}{p-1}-(s+1)}\epsilon^2 \log\frac{1}{\epsilon}\int_{R^2} w^r(y)\,dy \sim 1.$$

Hence we should look for solutions of (1.1) with the following properties:

$$A_\epsilon = \xi_\epsilon^{q/(p-1)}\bar{A}_\epsilon, \quad \bar{A}_\epsilon(y) = w(y) + \phi_\epsilon(y), \quad \phi_\epsilon \sim 0,$$

$$\text{where} \quad y = \frac{x-P_\epsilon}{\epsilon} \quad \text{and} \quad |P_\epsilon - P_0| = o(1) \ \text{ as } \ \epsilon \to 0,$$

$$H_\epsilon = \xi_\epsilon \bar{H}_\epsilon, \quad \bar{H}_\epsilon(x) = 1 + \psi_\epsilon(x), \quad \psi_\epsilon \sim 0,$$

and

$$\xi_\epsilon^{\frac{qr}{p-1}-(s+1)}\epsilon^2 \log\frac{1}{\epsilon}\int_{R^2} w^r(y)\,dy \sim 1.$$

There are three main difficulties: First, $w(\frac{x-P_\epsilon}{\epsilon})$ does not satisfy the Neumann boundary condition. Second, the linearized problem arising from (1.4) has the N-dimensional kernel $\text{span}\{\frac{\partial w}{\partial y_1}, \dots, \frac{\partial w}{\partial y_N}\}$. Therefore, if we linearize system (1.6) at $(w(\frac{x-P}{\epsilon}), 1)$, the linearized operator is not uniformly invertible with respect to ϵ. Third, we have two scales: $(\log\frac{1}{\epsilon})^{-1}$ and ϵ. They are simply incomparable.

The first difficulty can be overcome by introducing the following *projection*: Let $U \subset R^2$ be a smooth and open set. Suppose that $W \in H^1(R^2)$. The projection $\mathcal{P}_U W$ is defined by $\mathcal{P}_U W = W - \mathcal{Q}_U W$, where $\mathcal{Q}_U W$ satisfies

$$\begin{cases} \Delta \mathcal{Q}_U W - \mathcal{Q}_U W = 0 & \text{in } U, \\ \frac{\partial \mathcal{Q}_U W}{\partial \nu} = \frac{\partial W}{\partial \nu} & \text{on } \partial U. \end{cases} \tag{1.7}$$

The second difficulty is overcome by first "solving" (1.6) module approximate kernel and cokernel, respectively. Subsequently we use the nondegeneracy of the critical point of h at P_0 to choose P_ϵ near P_0 such that the finite-dimensional part lying in the approximate cokernel vanishes.

The third difficulty can be managed by choosing *suitable* approximate solutions.

From now on, we work with (1.6). The main points of the proof of Theorem 1.1 and the organization of this paper can be described as follows:

(A) Choose good approximate solutions.

We first study the solution $(A_{\epsilon,\mu}(x), H_{\epsilon,\mu}(x), c_{\epsilon,\mu})$ of the following problem:

$$\begin{cases} \epsilon^2 \Delta A - A + \frac{A^p}{(H(x)-\mu)^q} = 0, & x \in R^2, \\ \Delta H - H + c_{\epsilon,\mu} \frac{A^r}{(H(x)-\mu)^s} = 0, & x \in R^2, \\ H(0) = 1 + O(\frac{1}{\log \frac{1}{\epsilon}} + \mu), \end{cases} \tag{1.8}$$

where μ is small.

Next we choose $\mu := \mu_\epsilon(P)$ so that

$$\mu = \mathcal{Q}_\Omega(H_{\epsilon,\mu}(\cdot - P))(P). \tag{1.9}$$

Set

$$\hat{A}_{\epsilon,P}(x) := A_{\epsilon,\mu_\epsilon(P)}(x - P), \quad \hat{H}_{\epsilon,P}(x) := H_{\epsilon,\mu_\epsilon(P)}(x - P),$$

$$c_\epsilon = \xi^{\frac{qr}{p-1}-(s+1)}, \quad c_{\epsilon,P} := c_{\epsilon,\mu_\epsilon(P)}.$$

We now choose our approximate solutions:

$$A_{\epsilon,P}(y) := \mathcal{P}_{\Omega_{\epsilon,P}} \hat{A}_{\epsilon,P}(P + \epsilon y), \quad H_{\epsilon,P}(x) := \mathcal{P}_\Omega \hat{H}_{\epsilon,P}(x). \tag{1.10}$$

Set

$$\varphi_{\epsilon,P}(y) := \hat{A}_{\epsilon,P}(y) - A_{\epsilon,P}(y), \quad \psi_{\epsilon,P}(x) := \hat{H}_{\epsilon,P}(x) - H_{\epsilon,P}(x).$$

It will be proved that $\varphi_{\epsilon,P}(y) = O(e^{-d(P,\partial\Omega)/\epsilon})$ for almost everywhere (a.e.) $y \in \Omega_{\epsilon,P}$ and $\psi_{\epsilon,P} = \frac{1}{\log \frac{1}{\epsilon}}(H(P,x) + o(1))$ uniformly with respect to $x \in \Omega$.

We will analyze $A_{\epsilon,P}$ and $H_{\epsilon,P}$ in sections 2 and 3.

(B) The idea now is to look for a solution of (1.6) of the form

$$\bar{A}_\epsilon(P + \epsilon y) = A_{\epsilon,P}(y) + \phi(y), \quad \bar{H}_\epsilon(x) = H_{\epsilon,P}(x) + \psi(x).$$

We will show that, provided P is properly chosen, ϕ and ψ are expected to be insignificantly small.

We now write system (1.6) in operator form.

For any smooth and open set $U \subset R^2$, let

$$W_N^{2,t}(U) = \left\{ u \in W^{2,t}(U) | \frac{\partial u}{\partial \nu} = 0 \text{ on } \partial U \right\}, \quad H_N^2(U) = W_N^{2,2}(U).$$

For $A(y) \in H_N^2(\Omega_{\epsilon,P})$, $H(x) \in W_N^{2,t}(\Omega)$, where $1 < t < 1.1$. (We need $t > 1$ so that the Sobolev embedding $W^{2,t}(\Omega) \subset L^\infty(\Omega)$ is continuous.) Set

$$S_\epsilon \begin{pmatrix} A \\ H \end{pmatrix} = \begin{pmatrix} S_1(A,H) \\ S_2(A,H) \end{pmatrix},$$

where $S_1(A,H) = \Delta_y A - A + A^p/H^q$, $S_2(A,H) = \Delta_x H - H + c_{\epsilon,P} A^r/H^s$.

Then solving (1.6) is equivalent to

$$S_\epsilon \begin{pmatrix} A \\ H \end{pmatrix} = 0, \quad A \in H_N^2(\Omega_{\epsilon,P}), \quad H \in W_N^{2,t}(\Omega). \tag{1.11}$$

We now substitute $A = A_{\epsilon,P}(y) + \phi(y)$, $H = H_{\epsilon,P}(x) + \psi(x)$ into (1.11). The system determining ϕ and ψ can be written as

$$S'_\epsilon \begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P} \end{pmatrix} \begin{bmatrix} \phi \\ \psi \end{bmatrix} + \begin{pmatrix} E^1_{\epsilon,P} \\ E^2_{\epsilon,P} \end{pmatrix} + \begin{pmatrix} O(\|\phi\|^2_{L^2(\Omega_{\epsilon,P})} + \|\psi\|^2_{L^t(\Omega)}) \\ O(\|\phi\|^2_{L^2(\Omega_{\epsilon,P})} + \|\psi\|^2_{L_t(\Omega)}) \end{pmatrix} = 0,$$

where $E^i_{\epsilon,P}, i = 1, 2$ denote the error terms and $E^1_{\epsilon,P} = S_1(A_{\epsilon,P}, H_{\epsilon,P}), E^2_{\epsilon,P} = S_2(A_{\epsilon,P}, H_{\epsilon,P})$. We will estimate the error terms in section 3.

It is then natural to try to solve the equations for (ϕ, ψ) by a contraction mapping argument. The problem is that the linearized operator $S'_\epsilon(\begin{smallmatrix} A_{\epsilon,P} \\ H_{\epsilon,P} \end{smallmatrix})$ is not uniformly invertible with respect to ϵ.

Therefore, we now replace the above equation with

$$S'_\epsilon \begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P} \end{pmatrix} \begin{bmatrix} \phi \\ \psi \end{bmatrix} + \begin{pmatrix} E^1_{\epsilon,P} \\ E^2_{\epsilon,P} \end{pmatrix} + \begin{pmatrix} O(\|\phi\|^2_{L^2(\Omega_{\epsilon,P})} + \|\psi\|^2_{L^t(\Omega)}) \\ O(\|\phi\|^2_{L^2(\Omega_{\epsilon,P})} + \|\psi\|^2_{L_t(\Omega)}) \end{pmatrix} = \begin{pmatrix} v_{\epsilon,P} \\ 0 \end{pmatrix},$$

(1.12)

where $v_{\epsilon,P}$ lies in an appropriately chosen approximate cokernel of the linear operator

$$L_\epsilon := \Delta_y - 1 + pA^{p-1}_{\epsilon,P} H^{-q}_{\epsilon,P} - \frac{qr}{s+1} \frac{\int_{\Omega_{\epsilon,P}} A^{r-1}_{\epsilon,P} \cdot}{\int_{\Omega_{\epsilon,P}} A^r_{\epsilon,P}} A^p_{\epsilon,P},$$

$$L_\epsilon : H^2(\Omega_{\epsilon,P}) \to L^2(\Omega_{\epsilon,P}),$$

and ϕ is orthogonal in $L^2(\Omega_{\epsilon,P})$ to the corresponding approximate kernel of L_ϵ.

(C) We solve (1.12) for (ϕ, ψ) module the approximate kernel. To this end, we need a detailed analysis of the operators L_ϵ and S'_ϵ. This together with the contraction mapping argument is done in section 4.

(D) In the last step, we study a vector field $P \to W_\epsilon(P)$ such that $W_\epsilon(P) = 0$ implies $v_{\epsilon,P} = 0$ (and hence solutions of system (1.6) can be found). To discuss the zeros of $P \to W_\epsilon(P)$ we need very good estimates for the error terms $E^1_{\epsilon,P}$ and $E^2_{\epsilon,P}$. Much of section 3 is devoted to this analysis. With a good estimate of $E^i_{\epsilon,P}, i = 1, 2$, we discover that under the geometric condition described in Theorem 1.1 there is a point P_ϵ in a small neighborhood of $P_0 \in \Omega$ such that $W_\epsilon(P_\epsilon) = 0$. This will complete the proof of Theorem 1.1 and is done in section 5.

Finally, we remark that the stability of the solutions constructed in Theorem 1.1 should be related to the matrix $(\nabla_i \nabla_j h(P_0))$. This will be studied in a forthcoming paper.

Throughout this paper, we always assume that $P \in B_r(P_0)$ for some fixed small number $r > 0$. We shall frequently use the following technical lemma.

LEMMA 1.2. *Let u be a solution of*

$$\Delta u - u + f = 0 \quad \text{in } \Omega, \qquad \frac{\partial u}{\partial \nu} = 0 \quad \text{on } \partial\Omega.$$

Suppose

$$|f(x)| \le \eta e^{-\frac{\alpha|x-P|}{\epsilon}}$$

for some $\alpha > 0$. Then if $\epsilon > 0$ is small enough we have

$$|u(P)| \le C_1 \eta \epsilon^2 \log \frac{1}{\epsilon} \tag{1.13}$$

and

$$|u(P)-u(x)| \le C_2\eta\epsilon^2 \log\left(\frac{|x-P|}{\epsilon}+1\right), \tag{1.14}$$

where $C_1>0, C_2>0$ are generic constants (which are independent of $\epsilon>0$ and $\eta>0$).

Proof. By the representation formula we calculate

$$u(x)=\int_\Omega G(x,z)f(z)dz$$

and

$$u(P)=\int_\Omega G(P,z)f(z)dz=\epsilon^2\int_{\Omega_{\epsilon,P}} G(P,P+\epsilon y)f(P+\epsilon y)dy$$

$$\le C_1\eta\epsilon^2\log\frac{1}{\epsilon}.$$

Similarly we can obtain (1.14). ∎

2. Study of the approximate solutions. In this section, we define a good approximate solution and study its properties. We will use the implicit function theorem and perturbation arguments. To this end, it is essential that we have the following important lemma.

LEMMA 2.1. *The operator*

$$L:=\Delta-1+pw^{p-1}-\frac{qr}{s+1}\frac{\int_{R^2} w^{r-1}.}{\int_{R^2} w^r}w^p$$

with w defined in (1.4) *is an invertible map from $H_r^2(R^2)$ to $L_r^2(R^2)$, where $H_r^2(R^2)$ ($L_r^2(R^2)$) is the subset of those functions of $H^2(R^2)$ ($L_r^2(R^2)$) that are radially symmetric.*

Proof. We just need to prove that

$$\text{kernel}(L)\cap H_r^2(R^2)=\{0\},\quad \text{kernel}(L^*)\cap H_r^2(R^2)=\{0\},$$

where L^* is the conjugate operator of L.

In fact, let $L\phi=0$ for $\phi\in H_r^2(R^2)$. Then we have

$$L_0\left(\phi-\frac{qr}{(p-1)(s+1)}\frac{\int_{R^2} w^{r-1}\phi}{\int_{R^2} w^r}w\right)=0,$$

where $L_0:=\Delta-1+pw^{p-1}$. By Lemma 4.2 of [30], $\phi-\frac{qr}{(p-1)(s+1)}\frac{\int_{R^2} w^{r-1}\phi}{\int_{R^2} w^r}w=0$. Multiplying this equation by w^{r-1} and integrating over R^2 we see that

$$\int_{R^2} w^{r-1}\phi=0.$$

Since $\frac{qr}{(p-1)(s+1)}>1$ we conclude $\phi=0$.

Next we claim that $\text{kernel}(L^*)\cap H_r^2(R^2)=\{0\}$. Let $\phi\in H_r^2(R^2)$ be such that $L^*\phi=0$. Namely, we have

$$L_0\phi-\frac{qr}{s+1}\frac{\int_{R^2} w^p\phi}{\int_{R^2} w^r}w^{r-1}=0. \tag{2.1}$$

Multiplying (2.1) by w and integrating over R^2, we obtain

$$\left(p-1-\frac{qr}{s+1}\right)\int_{R^2} w^p\phi = 0.$$

Since $p-1-\frac{qr}{s+1} < 0$ we get

$$\int_{R^2} w^p\phi = 0.$$

Hence $L_0\phi = 0$ and $\phi = 0$. □

We now study the following system:

$$\begin{cases} \epsilon^2\Delta A - A + \frac{A^p}{(H-\mathcal{Q}_\Omega H(P))^q} = 0, \quad x\in R^2, \\ \Delta H - H + c_{\epsilon,P}\frac{A^r}{(H-\mathcal{Q}_\Omega H(P))^s} = 0, \quad x\in R^2, \\ H(P) = 1 + O(\frac{1}{\log\frac{1}{\epsilon}}). \end{cases} \tag{2.2}$$

We have Theorem 2.2.

THEOREM 2.2. *For $\epsilon << 1$, there exists a unique solution $(\hat A_{\epsilon,P}(x), \hat H_{\epsilon,P}(x), c_{\epsilon,P})$ of (2.2) with the following properties:*

(1) *$\hat A_{\epsilon,P}(x)$ and $\hat H_{\epsilon,P}(x)$ depend on $|x-P|$ only;*

(2) *$\hat A_{\epsilon,P} = (1+o(1))w(\frac{|x-P|}{\epsilon})$;*

(3) *$\hat H_{\epsilon,P}(0) = 1 + O(\frac{1}{\log\frac{1}{\epsilon}})$;*

(4) *$\hat H_{\epsilon,P}(x) = \frac{\sigma_P}{\log\frac{1}{\epsilon}}K(|x-P|) + \frac{\epsilon}{\log\frac{1}{\epsilon}}J_{\epsilon,P}(|x-P|)$ for $|x|\geq\delta$, where $\sigma_P = 1+o(1)$, $J_{\epsilon,P}(|x-P|)$, $\nabla_x J_{\epsilon,P}(|x-P|) = O(1)$.*

Proof of Theorem 2.2. The proof is divided into the following steps:

Step 1. We first look for radially symmetric solutions $(A_{\epsilon,\mu}, H_{\epsilon,\mu}, c_{\epsilon,\mu})$ of the following parametrized equation:

$$\begin{cases} \epsilon^2\Delta A - A + \frac{A^p}{(H-\mu)^q} = 0, \quad x\in R^2, \\ \Delta H - H + c_{\epsilon,\mu}\frac{A^r}{(H-\mu)^s} = 0, \quad x\in R^2, \\ A(x) = A(|x|), \quad H(x) = H(|x|), \quad H(0) = 1 + O(\frac{1}{\log\frac{1}{\epsilon}} + \mu) \end{cases} \tag{2.3}$$

for $0 < \mu << 1$.

Problem (2.3) can be solved by the contraction mapping principle. We first need suitable approximate solutions. We note that the problem

$$\begin{cases} \Delta_y A - A + \frac{A^p}{(1-\mu)^q} = 0, \quad y\in R^2, \\ \Delta_x H - H + c_{\epsilon,\mu,0}\frac{A^r}{(H-\mu)^s} = 0, \quad x\in R^2, \\ x = \epsilon y, A(y) = A(|y|), H(x) = H(|x|), H(0) = 1 \end{cases} \tag{2.4}$$

has a unique solution $(A_{\epsilon,\mu,0}(y), H_{\epsilon,\mu,0}(x), c_{\epsilon,\mu,0})$ for $0 < \mu << 1$. In fact, it is well known that (for given μ small) the first equation has the unique positive solution $A_{\epsilon,\mu,0}(y) = (1-\mu)^{q/(p-1)}w(y)$ with maximum at 0 and decaying to 0 at infinity (compare equation (1.4)). It is also easy to see that for given μ and $A\in H^2(R^2)$, the second equation has a unique solution $H_{\epsilon,\mu,0}(x)\in H^2(R^2)$ (note that the nonlinearity is concave). To ensure that $H_{\epsilon,\mu,0}(0) = 1$, we just need to choose $c_{\epsilon,\mu,0}$. In fact, by the standard representation formula

$$H_{\epsilon,\mu,0}(x) = \int_{R^2} K(|x-z|)\, c_{\epsilon,\mu,0}\,(1-\mu)^{rq/(p-1)}(H_{\epsilon,\mu,0}-\mu)^{-s}(z)w^r\left(\frac{z}{\epsilon}\right)\,dz.$$

Taking $x = 0$, we obtain

$$c_{\epsilon,\mu,0} = (1-\mu)^{s-rq/(p-1)} \left(\int_{R^2} K(|z|) \left(1 + O\left(\frac{1}{\log \frac{1}{\epsilon}} + \mu \right) \right) w^r \left(\frac{z}{\epsilon} \right) \, dz \right)^{-1}$$
$$= (1-\mu)^{s-rq/(p-1)} \left(\epsilon^2 \left(1 + O\left(\left(\log \frac{1}{\epsilon} \right)^{-1} \right) + \mu \right) \int_{R^2} K(|\epsilon y|) w^r(y) \, dy \right)^{-1}$$
$$= (1-\mu)^{s-rq/(p-1)} \frac{1}{\epsilon^2 \log(1/\epsilon)} \left(\int_{R^2} w^r(y) \, dy \right)^{-1} + O\left(\frac{1/\log(1/\epsilon) + \mu}{\epsilon^2 \log(1/\epsilon) 2} \right)$$

as $\epsilon \to 0$.

(Here we have used the fact that (by Lemma 1.2)

$$|H_{\epsilon,\mu,0}(x) - H_{\epsilon,\mu,0}(0)| \le C \frac{1}{\log \frac{1}{\epsilon}} \log \left(\frac{|x-P|}{\epsilon} + 1 \right)$$

for some generic constant $C > 0$.)

Using the ansatz

$$A_{\epsilon,\mu}(y) = A_{\epsilon,\mu,0}(y) + a_{\epsilon,\mu}(y),$$

$$H_{\epsilon,\mu}(x) = H_{\epsilon,\mu,0}(x) + h_{\epsilon,\mu}(x),$$

and inserting it into (2.3) (with $c_{\epsilon,\mu} = c_{\epsilon,\mu,0}$) gives us

$$\Delta_y a_{\epsilon,\mu} - a_{\epsilon,\mu} = \frac{A^p_{\epsilon,\mu,0}}{(1-\mu)^q} - \frac{(A_{\epsilon,\mu,0} + a_{\epsilon,\mu})^p}{(H_{\epsilon,\mu,0} + h_{\epsilon,\mu} - \mu)^q},$$

$$\Delta_x h_{\epsilon,\mu} - h_{\epsilon,\mu} = c_{\epsilon,\mu,0} \frac{A^r_{\epsilon,\mu,0}}{(H_{\epsilon,\mu,0} - \mu)^s} - c_{\epsilon,\mu,0} \frac{(A_{\epsilon,\mu,0} + a_{\epsilon,\mu})^r}{(H_{\epsilon,\mu,0} + h_{\epsilon,\mu} - \mu)^s}.$$

The first equation can be rewritten as follows:

$$\Delta_y a_{\epsilon,\mu} - a_{\epsilon,\mu} + \frac{p A^{p-1}_{\epsilon,\mu,0} a_{\epsilon,\mu}}{(1-\mu)^q} - \frac{q A^p_{\epsilon,\mu,0} h_{\epsilon,\mu}}{(1-\mu)^{q+1}} = e_1,$$

where

$$e_1 = \frac{(A_{\epsilon,\mu,0} + a_{\epsilon,\mu})^p}{(1 + h_{\epsilon,\mu} - \mu)^q} - \frac{(A_{\epsilon,\mu,0} + a_{\epsilon,\mu})^p}{(H_{\epsilon,\mu,0} + h_{\epsilon,\mu} - \mu)^q}$$
$$- \frac{(A_{\epsilon,\mu,0} + a_{\epsilon,\mu})^p}{(1 + h_{\epsilon,\mu} - \mu)^q} + \frac{A^p_{\epsilon,\mu,0}}{(1-\mu)^q} + \frac{p A^{p-1}_{\epsilon,\mu,0} a_{\epsilon,\mu}}{(1-\mu)^q} - \frac{q A^p_{\epsilon,\mu,0} h_{\epsilon,\mu}}{(1-\mu)^{q+1}}.$$

This implies

$$\|e_1(y)\|_{L^2(R^2)} = O\left(\|a_{\epsilon,\mu}(y)\|^2_{L^2(R^2)} \right) + O(\|h_{\epsilon,\mu}(x)\|^2_{L^\infty(\Omega)}) + O\left(\frac{1}{\log \frac{1}{\epsilon}} + \mu \right).$$

For a given $a_{\epsilon,\mu}$, we can solve the second equation directly since the nonlinearity is concave. Moreover, we have that $h_{\epsilon,\mu}$ satisfies

$$\Delta_x h_{\epsilon,\mu} - h_{\epsilon,\mu} + c_{\epsilon,\mu} \frac{r A^{r-1}_{\epsilon,\mu,0} a_{\epsilon,\mu}}{(H_{\epsilon,\mu,0} - \mu)^s} - c_{\epsilon,\mu} \frac{s A^r_{\epsilon,\mu,0} h_{\epsilon,\mu}}{(H_{\epsilon,\mu,0} - \mu)^{s+1}} = e_2,$$

where

$$e_2 = c_{\epsilon,\mu}\frac{A^r_{\epsilon,\mu,0}}{(H_{\epsilon,\mu,0}-\mu)^s} - c_{\epsilon,\mu}\frac{(A_{\epsilon,\mu,0}+a_{\epsilon,\mu})^r}{(H_{\epsilon,\mu,0}-\mu)^s}$$

$$+c_{\epsilon,\mu}\frac{rA^{r-1}_{\epsilon,\mu,0}a_{\epsilon,\mu}}{(H_{\epsilon,\mu,0}-\mu)^s} - c_{\epsilon,\mu}\frac{sA^r_{\epsilon,\mu,0}h_{\epsilon,\mu}}{(H_{\epsilon,\mu,0}-\mu)^{s+1}}.$$

This implies

$$\|e_2\|_{L^2(R^2)} = O(\|a_{\epsilon,\mu}\|^2_{L^2(R^2)}) + O(\|h_{\epsilon,\mu}\|^2_{L^\infty(\Omega)}\|A^{r-1}_{\epsilon,\mu,0}\|_{L^2(R^2)}).$$

Thus by Lemma 1.2

$$h_{\epsilon,\mu}(x) = h_{\epsilon,\mu}(0) + O\left(\frac{1}{\log\frac{1}{\epsilon}}\right)$$

and

$$h_{\epsilon,\mu}(0) = \int_{R^2} K(z)\left[c_{\epsilon,\mu}\frac{rA^{r-1}_{\epsilon,\mu,0}a_{\epsilon,\mu}}{(H_{\epsilon,\mu,0}-\mu)^s} - c_{\epsilon,\mu}\frac{sA^r_{\epsilon,\mu,0}h_{\epsilon,\mu}}{(H_{\epsilon,\mu,0}-\mu)^{s+1}}\right] + O(\|a_{\epsilon,\mu}\|^2_{L^2(R^2)})$$

$$= c_{\epsilon,\mu}\int_{R^2} rA^{r-1}_{\epsilon,\mu,0}a_{\epsilon,\mu}\left(1+O\left(\frac{1}{\log\frac{1}{\epsilon}}+\mu\right)\right)$$

$$-c_{\epsilon,\mu}sh_{\epsilon,\mu}(0)\int_{R^2} A^r_{\epsilon,\mu,0}\left(1+O\left(\frac{1}{\log\frac{1}{\epsilon}}+\mu\right)\right) + O(\|a_{\epsilon,\mu}\|^2_{L^2(R^2)}).$$

Therefore

$$h_{\epsilon,\mu}(0) = \frac{r}{s+1}\frac{\int A^{r-1}_{\epsilon,\mu,0}a_{\epsilon,\mu}}{\int A^r_{\epsilon,\mu,0}} + O\left(\frac{1}{\log\frac{1}{\epsilon}}+\mu\right) + O(\|a_{\epsilon,\mu}\|^2_{L^2(R^2)}).$$

Substituting this into the first equation, the equation for $a_{\epsilon,\mu}$ becomes

$$\Delta_y a_{\epsilon,\mu} - a_{\epsilon,\mu} + \frac{pA^{p-1}_{\epsilon,\mu,0}a_{\epsilon,\mu}}{(1-\mu)^q} - \frac{qrA^p_{\epsilon,\mu,0}}{(s+1)(1-\mu)^{q+1}}\frac{\int_{R^2} A^{r-1}_{\epsilon,\mu,0}a_{\epsilon,\mu}}{\int_{R^2} A^r_{\epsilon,\mu,0}}$$

$$= e_1 + O\left(\frac{1}{\log\frac{1}{\epsilon}}+\mu\right) + O(\|a_{\epsilon,\mu}\|^2_{L^2(R^2)})$$

in $L^2(R^2)$.

By Lemma 2.1 and a perturbation argument for $\epsilon << 1, \mu << 1$, the equation for $a_{\epsilon,\mu}$ can be solved and the solution is unique. Thus we have obtained a solution to (2.3).

Step 2. We choose μ such that

$$\mu = H_{\epsilon,\mu}(0) - \mathcal{P}_\Omega(H_{\epsilon,\mu}(\cdot - P))(P). \tag{2.5}$$

To this end, we note that this is equivalent to

$$\mu = \int_{R^2}(K(|z|) - G(P,P+z))\,c_{\epsilon,\mu}\,(H_{\epsilon,\mu}(z)-\mu)^{-s}A^r_{\epsilon,\mu}\left(\frac{z}{\epsilon}\right)\,dz$$

$$= \int_{R^2} H(P, P+z) c_{\epsilon,\mu} (H_{\epsilon,\mu}(z) - \mu)^{-s} A^r_{\epsilon,\mu}\left(\frac{z}{\epsilon}\right) dz$$

$$= H(P,P) c_{\epsilon,\mu} \int_{R^2} (H_{\epsilon,\mu}(z) - \mu)^{-s} A^r_{\epsilon,\mu}\left(\frac{z}{\epsilon}\right) dz$$

$$+ O(\epsilon) \int_{R^2} |z| c_{\epsilon,\mu} (H_{\epsilon,\mu}(z) - \mu)^{-s} A^r_{\epsilon,\mu}\left(\frac{z}{\epsilon}\right) dz.$$

Since $c_{\epsilon,\mu} \int_{R^2} (H_{\epsilon,\mu}(z) - \mu)^{-s} A^r_{\epsilon,\mu}(\frac{z}{\epsilon}) dz = \frac{1+o(1)}{\log \frac{1}{\epsilon}}$, it is easy to see that by the contraction mapping principle, (1.9) has a unique solution $\mu = \mu_\epsilon(P)$.

We further calculate

$$\mu = \frac{1+o(1)}{\log \frac{1}{\epsilon}} \left[H(P,P) + O\left(\frac{1}{\log \frac{1}{\epsilon}}\right)\right]$$

as $\epsilon \to 0$.

Now let

$$\hat{A}_{\epsilon,P}(x) = A_{\epsilon,\mu}(x-P), \quad \hat{H}_{\epsilon,P}(x) = H_{\epsilon,\mu}(x-P), \quad c_{\epsilon,P} = c_{\epsilon,\mu},$$

where $\mu := \mu_\epsilon(P)$ is given by (1.9).

It is easy to see that (1), (2), and (3) of Theorem 1.1 are satisfied. It remains to prove statement (4) of Theorem 2.2. We have for $|x| \geq \delta$,

$$\hat{H}_{\epsilon,P}(x) = \frac{\epsilon^2 \int_{R^2} K(|x-P-\epsilon y|) \frac{A^r_{\epsilon,\mu}(y)}{H^s_{\epsilon,\mu}} dy}{\int_{R^2} K(|\epsilon y|) \frac{A^r_{\epsilon,\mu}(y)}{H^s_{\epsilon,\mu}} dy}$$

$$= \frac{\sigma_P}{\log \frac{1}{\epsilon}} [K(|x-P|) + O(\epsilon)], \quad \sigma_P = 1 + o(1)$$

as $\epsilon \to 0$.

This implies Theorem 2.2. □

3. Estimates of the error terms. In this section, we give some preliminary estimates. These will be used in the later sections.

Recall that we choose our approximate solution as follows:

$$A_{\epsilon,P}(y) = \mathcal{P}_{\Omega_{\epsilon,P}} \hat{A}_{\epsilon,P}, \quad H_{\epsilon,P}(x) = \mathcal{P}_\Omega \hat{H}_{\epsilon,P}(x).$$

Note that in this case

$$\mu = \mathcal{Q}_\Omega \hat{H}_{\epsilon,P}(P).$$

Also recall that

$$\varphi_{\epsilon,P}(y) = \mathcal{Q}_{\Omega_{\epsilon,P}} \hat{A}_{\epsilon,P} = \hat{A}_{\epsilon,P} - A_{\epsilon,P}, \quad \psi_{\epsilon,P}(x) = \mathcal{Q}_\Omega \hat{H}_{\epsilon,P} = \hat{H}_{\epsilon,P} - H_{\epsilon,P}.$$

We note that $\varphi_{\epsilon,P}$ satisfies

$$\Delta_y \varphi_{\epsilon,P} - \varphi_{\epsilon,P} = 0 \quad \text{in } \Omega_{\epsilon,P},$$

$$\frac{\partial \varphi_{\epsilon,P}}{\partial \nu} = \frac{\partial \hat{A}_{\epsilon,P}}{\partial \nu} = O(e^{-d(P,\partial\Omega)/\epsilon}) \quad \text{on } \partial\Omega_{\epsilon,P}.$$

Hence

$$\|\varphi_{\epsilon,P}\|_{H^2(\Omega_{\epsilon,P})} = O(e^{-d(P,\partial\Omega)/\epsilon}). \tag{3.1}$$

By Theorem 2.2 we have

$$\mathcal{P}_\Omega \hat{H}_{\epsilon,P}(x) = \frac{\int_{\Omega_{\epsilon,P}} G(x, P+\epsilon y) \frac{\hat{A}^r_{\epsilon,P}(y)}{(\hat{H}_{\epsilon,P}-\mu_\epsilon(P))^s}\, dy}{\int_{R^2} K(|\epsilon y|) \frac{\hat{A}^r_{\epsilon,P}(y)}{(\hat{H}_{\epsilon,P}-\mu_\epsilon(P))^s}\, dy}$$

$$= \frac{1+o(1)}{\log\frac{1}{\epsilon}}[K(|x-P|) - H(x,P) + O(\epsilon)].$$

This implies

$$\psi_{\epsilon,P}(x) = \hat{H}_{\epsilon,P}(x) - \mathcal{P}_\Omega \hat{H}_{\epsilon,P}(x-P) = \frac{1+o(1)}{\log\frac{1}{\epsilon}}[H(x,P) + O(\epsilon)]$$

or, equivalently,

$$\psi_{\epsilon,P}(x) = \frac{1+o(1)}{\log\frac{1}{\epsilon}} H(P,x) + O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right). \tag{3.2}$$

By (3.1) and (3.2), we see that the term involving $\varphi_{\epsilon,P}$ can be neglected. This is what we will do in the later sections.

The reason for choosing $A_{\epsilon,\mu}$ and $H_{\epsilon,P}$ as we did lies in the two following estimates:

$$S_1(A_{\epsilon,P}, H_{\epsilon,P}) = \Delta_y A_{\epsilon,P} - A_{\epsilon,P} + \frac{A^p_{\epsilon,P}}{H^q_{\epsilon,P}}$$

$$= \frac{(\hat{A}_{\epsilon,P}-\varphi_{\epsilon,P})^p}{(\hat{H}_{\epsilon,P}-\psi_{\epsilon,P})^q} - \frac{(\hat{A}_{\epsilon,P})^p}{(\hat{H}_{\epsilon,P}-\psi_{\epsilon,P}(P))^q}$$

$$= O(e^{-d(P,\partial\Omega)/\epsilon}) + (\hat{A}_{\epsilon,P})^p[(\hat{H}_{\epsilon,P}-\psi_{\epsilon,P})^{-q} - (\hat{H}_{\epsilon,P}-\psi_{\epsilon,P}(P))^{-q}] \quad \text{(by (3.1))}$$

$$= O(e^{-d(P,\partial\Omega)/\epsilon}) - q(\hat{A}_{\epsilon,P})^p(\hat{H}_{\epsilon,P})^{-q-1}(\psi_{\epsilon,P}(x)-\psi_{\epsilon,P}(P)) + O\left(\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right)^2 \hat{A}^p_{\epsilon,P}\right)$$

for a.e. $y \in \Omega_{\epsilon,P}$. Similarly we have

$$S_2(A_{\epsilon,P}, H_{\epsilon,P}) = \Delta_x H_{\epsilon,P} - H_{\epsilon,P} + c_{\epsilon,P}\frac{A^r_{\epsilon,P}}{H^s_{\epsilon,P}}$$

$$= O(e^{-d(P,\partial\Omega)/\epsilon}) - sc_{\epsilon,P}(\hat{A}_{\epsilon,P})^r(\hat{H}_{\epsilon,P})^{-s-1}(\psi_{\epsilon,P}(x)-\psi_{\epsilon,P}(P)) + O\left(c_{\epsilon,P}\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right)^2 \hat{A}^r_{\epsilon,P}\right)$$

for a.e. $x \in \Omega$.

We have thus obtained Lemma 3.1.

LEMMA 3.1. *We have*

$$\begin{aligned} S_1(A_{\epsilon,P}, H_{\epsilon,P}) &= O(e^{-d(P,\partial\Omega)/\epsilon}) - q(\hat{A}_{\epsilon,P})^p(\hat{H}_{\epsilon,P})^{-q-1}(\psi_{\epsilon,P}(x)-\psi_{\epsilon,P}(P)) \\ &+ O\left(\left(\tfrac{\epsilon}{\log\frac{1}{\epsilon}}\right)^2 \hat{A}^p_{\epsilon,P}\right) \end{aligned} \tag{3.3}$$

for a.e. $y \in \Omega_{\epsilon,P}$.

$$S_2(A_{\epsilon,P}, H_{\epsilon,P})$$

$$= O(e^{-d(P,\partial\Omega)/\epsilon}) - sc_{\epsilon,P}(\hat{A}_{\epsilon,P})^r(\hat{H}_{\epsilon,P})^{-s-1}(\psi_{\epsilon,P}(x) - \psi_{\epsilon,P}(P)) + O\left(c_{\epsilon,P}\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right)^2 \hat{A}^r_{\epsilon,P}\right)$$

(3.4)

for a.e. $x \in \Omega$.

Hence

$$\|S_1(A_{\epsilon,P}, H_{\epsilon,P})\|_{L^2(\Omega_{\epsilon,P})} = O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right), \tag{3.5}$$

$$\|S_2(A_{\epsilon,P}, H_{\epsilon,P})\|_{L^t(\Omega)} = O\left(\epsilon^{2t^{-1}-1}\left(\frac{1}{\log\frac{1}{\epsilon}}\right)^2\right) \tag{3.6}$$

for any $1 < t < 1.1$.

Proof. The lemma is proved by direct computation. □

4. The Liapunov–Schmidt reduction method. This section is devoted to studying the linearized operator defined by

$$\tilde{L}_{\epsilon,P} := S'_\epsilon \begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P} \end{pmatrix},$$

$$\tilde{L}_{\epsilon,P} : H^2_N(\Omega_{\epsilon,P}) \times W^{2,t}_N(\Omega) \to L^2(\Omega_{\epsilon,P}) \times L^t(\Omega),$$

where $1 < t < 1.1$ is a fixed number.

Set

$$K_{\epsilon,P} := \text{span}\left\{\frac{\partial A_{\epsilon,P}}{\partial P_j}\middle| j = 1, \dots, N\right\} \subset H^2_N(\Omega_{\epsilon,P}),$$

$$C_{\epsilon,P} := \text{span}\left\{\frac{\partial A_{\epsilon,P}}{\partial P_j}\middle| j = 1, \dots, N\right\} \subset L^2(\Omega_{\epsilon,P}),$$

$$L_\epsilon := \Delta - 1 + pA^{p-1}_{\epsilon,P}H^{-q}_{\epsilon,P} - \frac{qr}{s+1}\frac{\int_{\Omega_{\epsilon,P}} A^{r-1}_{\epsilon,P}\cdot}{\int_{\Omega_{\epsilon,P}} A^r_{\epsilon,P}}A^p_{\epsilon,P},$$

and

$$L_{\epsilon,P} := \pi_{\epsilon,P} \circ L_\epsilon : K^\perp_{\epsilon,P} \to C^\perp_{\epsilon,P},$$

where $\pi_{\epsilon,P}$ is the projection in $L^2(\Omega_{\epsilon,P})$ onto $C^\perp_{\epsilon,P}$.

We remark that since $A_{\epsilon,P}(y) = (1 + O(\frac{1}{\log\frac{1}{\epsilon}}))w(y)$, it is easy to see that

$$l_{\epsilon,P} := \pi_{\epsilon,P} \circ (\Delta - 1 + pA^{p-1}_{\epsilon,P}) : K^\perp_{\epsilon,P} \to C^\perp_{\epsilon,P}$$

is a one to one and surjective map. For the proof please see the proof of Propositions 6.1–6.2 in [41].

The following proposition is the key estimate in applying the Liapunov–Schmidt reduction method.

PROPOSITION 4.1. *For ϵ sufficiently small, the map $L_{\epsilon,P}$ is a one-to-one and surjective map. Moreover the inverse of $L_{\epsilon,P}$ exists and is bounded uniformly with respect to ϵ.*

Proof. We will follow the method used in [11], [35], [36], [41], and [44]. We first show that there exist constants $C > 0, \bar{\epsilon} > 0$ such that for all $\epsilon \in (0, \bar{\epsilon})$,

$$\|L_{\epsilon,P}\Phi\|_{L^2(\Omega_{\epsilon,P})} \geq C\|\Phi\|_{H^2(\Omega_{\epsilon,P})} \tag{4.1}$$

for all $\Phi \in K^{\perp}_{\epsilon,P}$.

Suppose that (4.1) is false. Then there exist sequences $\{\epsilon_k\}$, $\{P_k\}$, and $\{\phi_k\}$ with $P_k \in \Omega$, $\phi_k \in K^{\perp}_{\epsilon_k,P_k}$ such that

$$\|L_{\epsilon_k,P_k}\phi_k\|_{L^2(\Omega_{\epsilon_k,P_k})} \to 0, \tag{4.2}$$

$$\|\phi_k\|_{H^2(\Omega_{\epsilon_k,P_k})} = 1, \qquad k = 1, 2, \ldots . \tag{4.3}$$

Namely, we have the following situation:

$$\Delta_y\phi_k - \phi_k + pA^{p-1}_{\epsilon_k,P_k}H^{-q}_{\epsilon_k,P_k}\phi_k - \frac{qr}{s+1}\frac{\int_{\Omega_{\epsilon_k,P_k}} A^{r-1}_{\epsilon_k,P_k}\phi_k}{\int_{\Omega_{\epsilon_k,P_k}} A^{r}_{\epsilon_k,P_k}}A^{p}_{\epsilon_k,P_k} = f_k, \tag{4.4}$$

where

$$\|f_k\|_{L^2(\Omega_{\epsilon_k,P_k})} \to 0,$$

$$\phi_k \in K^{\perp}_{\epsilon_k,P_k}, \quad \|\phi_k\|_{H^2(\Omega_{\epsilon_k,P_k})} = 1. \tag{4.5}$$

We now show that this is impossible. Set $A_k = A_{\epsilon_k,P_k}, \Omega_k = \Omega_{\epsilon_k,P_k}$.

Note that

$$H_{\epsilon_k,P_k} = 1 + o(1) \ \text{ in } L^{\infty}(\Omega),$$

$$(\Delta_y - 1 + pA^{p-1}_k)\frac{A_k}{p-1} = A^p_k + o(1) \text{ in } L^2(\Omega_k).$$

Thus we have

$$(\Delta_y - 1 + pA^{p-1}_k)\left(\phi_k - \frac{qr}{(s+1)(p-1)}\frac{\int_{\Omega_k} A^{r-1}_k\phi_k}{\int_{\Omega_k} A^r_k}A_k\right) = f_k + o(1) \text{ in } L^2(\Omega_k).$$

Since the projection of A_k into K_{ϵ_k,P_k} is $o(1)$ in $H^2(\Omega_k)$ and the operator

$$\Delta_y - 1 + pA^{p-1}_k$$

is a one-to-one and invertible map (with the inverse bounded uniformly with respect to ϵ) from $K^{\perp}_{\epsilon_k,P_k}$ to $C^{\perp}_{\epsilon_k,P_k}$, we have

$$\phi_k - \frac{qr}{(s+1)(p-1)}\frac{\int_{\Omega_k} A^{r-1}_k\phi_k}{\int_{\Omega_k} A^r_k}A_k = o(1) \text{ in } H^2(\Omega_k). \tag{4.6}$$

Since $\frac{qr}{(p-1)(s+1)} > 1$, (4.6) implies that

$$\|\phi_k\|_{H^2(\Omega_k)} = o(1).$$

This is a contradiction!

Thus (4.1) holds and $L_{\epsilon,P}$ is a one-to-one map.

Next we show that $L_{\epsilon,P}$ is also surjective. To this end, we just need to show that the conjugate of $L_{\epsilon,P}$ (denoted by $L^*_{\epsilon,P}$) is injective from $K^\perp_{\epsilon,P}$ to $C^\perp_{\epsilon,P}$.

Let $L^*_{\epsilon,P}\phi \in C^\perp_{\epsilon,P}$, $\phi \in K^\perp_{\epsilon,P}$. Namely, we have

$$\Delta_y\phi - \phi + pA^{p-1}_{\epsilon,P}H^{-q}_{\epsilon,P}\phi - \frac{qr}{s+1}\frac{\int_{\Omega_{\epsilon,P}} A^p_{\epsilon,P}\phi}{\int_{\Omega_{\epsilon,P}} A^r_{\epsilon,P}}A^{r-1}_{\epsilon,P} \in C_{\epsilon,P}. \tag{4.7}$$

We can assume that $\|\phi\|_{H^2(\Omega_{\epsilon,P})} = 1$.

Multiplying (4.7) by $A_{\epsilon,P}$ and integrating over $\Omega_{\epsilon,P}$, we obtain

$$\left(p - 1 - \frac{qr}{s+1}\right)\int_{\Omega_{\epsilon,P}} A^p_{\epsilon,P}\phi = o(1)$$

or, equivalently,

$$\int_{\Omega_{\epsilon,P}} A^p_{\epsilon,P}\phi = o(1).$$

Hence ϕ satisfies

$$\Delta_y\phi - \phi + pA^{p-1}_{\epsilon,P}H^{-q}_{\epsilon,P}\phi + o(1) \in C_{\epsilon,P}, \quad \phi \in K^\perp_{\epsilon,P},$$

which implies that $\|\phi\|_{H^2(\Omega_{\epsilon,P})} = o(1)$. This is a contradiction!

Therefore, $L_{\epsilon,P}$ is also surjective. □

We now deal with system (1.6).

$\tilde{L}_{\epsilon,P}$ is not uniformly invertible in ϵ due to the approximate kernel

$$\mathcal{K}_{\epsilon,P} := K_{\epsilon,P} \oplus \{0\} \subset H^2_N(\Omega_{\epsilon,P}) \times W^{2,t}_N(\Omega).$$

We choose the approximate cokernel as follows:

$$\mathcal{C}_{\epsilon,P} := C_{\epsilon,P} \oplus \{0\} \subset L^2(\Omega_{\epsilon,P}) \times L^t(\Omega).$$

We then define

$$\mathcal{K}^\perp_{\epsilon,P} := K^\perp_{\epsilon,P} \oplus W^{2,t}_N(\Omega) \subset H^2_N(\Omega_{\epsilon,P}) \times W^{2,t}_N(\Omega),$$
$$\mathcal{C}^\perp_{\epsilon,P} := C^\perp_{\epsilon,P} \oplus L^t(\Omega) \subset L^2(\Omega_{\epsilon,P}) \times L^t(\Omega).$$

Let $\pi_{\epsilon,P}$ denote the projection in $L^2(\Omega_{\epsilon,P}) \times L^t(\Omega)$ onto $\mathcal{C}^\perp_{\epsilon,P}$. (Here the projection in the second component is the identity map.) We then show that the equation

$$\pi_{\epsilon,P} \circ S_\epsilon \begin{pmatrix} A_{\epsilon,P} + \Phi_{\epsilon,P} \\ H_{\epsilon,P} + \Psi_{\epsilon,P} \end{pmatrix} = 0$$

has the unique solution

$$\Sigma_{\epsilon,P} = \begin{pmatrix} \Phi_{\epsilon,P}(y) \\ \Psi_{\epsilon,P}(x) \end{pmatrix} \in \mathcal{K}^\perp_{\epsilon,P}$$

if ϵ is small enough.

As a preparation in the following two propositions we show the invertibility of the corresponding linearized operator.

PROPOSITION 4.2. *Let $\mathcal{L}_{\epsilon,P} = \pi_{\epsilon,P} \circ \tilde{L}_{\epsilon,P}$. There exist positive constants $\overline{\epsilon}, \lambda$ such that for all $\epsilon \in (0, \overline{\epsilon})$,*

$$\|\mathcal{L}_{\epsilon,P}\Sigma\|_{L^2(\Omega_{\epsilon,P})\times L^t(\Omega)} \geq \lambda \|\Sigma\|_{H^2(\Omega_{\epsilon,P})\times W^{2,t}(\Omega)} \tag{4.8}$$

for all $\Sigma \in \mathcal{K}^{\perp}_{\epsilon,P}$.

PROPOSITION 4.3. *There exists a positive constant $\overline{\overline{\epsilon}}$ such that for all $\epsilon \in (0, \overline{\overline{\epsilon}})$, the map*

$$\mathcal{L}_{\epsilon,P} = \pi_{\epsilon,P} \circ \tilde{L}_{\epsilon} : \mathcal{K}^{\perp}_{\epsilon,P} \to \mathcal{C}^{\perp}_{\epsilon,P}$$

is surjective.

Proof of Proposition 4.2. This proposition follows from Proposition 4.1. In fact, suppose that (4.8) is false. Then there exist sequences $\{\epsilon_k\}$, $\{P_k\}$, and $\{\Sigma_k\}$ with

$$P_k \in \Omega, \qquad \Sigma_k = \begin{pmatrix} \phi_k(y) \\ \psi_k(x) \end{pmatrix} \in \mathcal{K}^{\perp}_{\epsilon_k,P_k}$$

such that

$$\|\mathcal{L}_{\epsilon_k,P_k}\Sigma_k\|_{L^2(\Omega_{\epsilon_k,P_k})\times L^t(\Omega)} \to 0, \tag{4.9}$$

$$\|\Sigma_k\|_{H^2(\Omega_{\epsilon_k,P_k})\times W^{2,t}(\Omega)} = 1, \qquad k = 1, 2, \ldots . \tag{4.10}$$

Namely, we have the following situation:

$$\Delta_y\phi_k - \phi_k + pA^{p-1}_{\epsilon_k,P_k}H^{-q}_{\epsilon_k,P_k}\phi_k - qA^{p}_{\epsilon_k,P_k}H^{-q-1}_{\epsilon_k,P_k}\psi_k = f_k, \quad \|f_k\|_{L^2(\Omega_{\epsilon_k,P_k})} \to 0, \tag{4.11}$$

$$\Delta_x\psi_k - \psi_k + rc_{\epsilon_k,P_k}A^{r-1}_{\epsilon_k,P_k}H^{-s}_{\epsilon_k,P_k}\phi_k - sc_{\epsilon_k,P_k}A^{r}_{\epsilon_k,P_k}H^{-s-1}_{\epsilon_k,P_k}\psi_k = g_k, \tag{4.12}$$

where

$$\|g_k\|_{L^t(\Omega)} \to 0,$$

$$\phi_k \in K^{\perp}_{\epsilon_k,P_k}, \tag{4.13}$$

$$\|\phi_k\|^2_{H^2(\Omega_{\epsilon_k,P_k})} + \|\psi_k\|^2_{W^{2,t}(\Omega)} = 1. \tag{4.14}$$

We now show that this is impossible. Set $A_k = A_{\epsilon_k,P_k}, \Omega_k = \Omega_{\epsilon_k,P_k}$.

We first note that by (4.12) we have

$$\|\psi_k\|_{L^\infty(\Omega)} \leq C$$

and hence by Lemma 1.2 and Sobolev embedding,

$$|\psi_k(x) - \psi_k(P_k)| \leq C|x - P_k|^\alpha + \frac{1}{\log\frac{1}{\epsilon}} \log\left(\frac{|x-P|}{\epsilon} + 1\right)$$

for some $\alpha > 0$ since $t > 1$. Thus

$$\|A^p_k(\psi_k - \psi_k(P_k))\|_{L^2(\Omega_k)} \to 0, \quad k = 1, 2, \ldots \quad \text{in} \quad L^2(\Omega_k). \tag{4.15}$$

Moreover by (4.12),

$$\psi_k(P_k) = \int_{\Omega_k} G(P,z)(rc_{\epsilon_k,P_k}A^{r-1}_k H^{-s}_{\epsilon_k,P_k}\phi_k - sc_{\epsilon_k,P_k}A^{r}_k H^{-s-1}_{\epsilon_k,P_k}\psi_k - g_k)$$

$$= (1+o(1))rc_{\epsilon_k,P_k}\log\frac{1}{\epsilon_k}\int_{\Omega_k}A^{r-1}_{\epsilon_k,P_k}\phi_k - (1+o(1))s\psi_k(P_k)c_{\epsilon_k,P_k}\int_{\Omega_k}A^r_k + o(1).$$

Therefore

$$\psi_k(P_k) = \frac{r}{s+1}\frac{\int_{\Omega_k}A^{r-1}_k\phi_k}{\int_{\Omega_k}A^r_k} + o(1).$$

Thus we have

$$L_{\epsilon_k,P_k}\phi_k = o(1) \quad \text{in} \quad L^2(\Omega_k), \quad \phi_k \in K^{\perp}_{\epsilon_k,P_k}. \tag{4.16}$$

By Proposition 4.1, $\|\phi_k\|_{H^2(\Omega_k)} = o(1)$. Hence $\psi_k(P_k) = o(1)$ and by elliptic estimates $\|\psi_k\|_{W^{2,t}(\Omega)} = o(1)$.

This contradicts assumption (4.14) and the proof of Proposition 4.2 is complete. □

Proof of Proposition 4.3. We just need to show that the conjugate operator of $\mathcal{L}_{\epsilon,P}$ (denoted by $\mathcal{L}^*_{\epsilon,P}$) is injective from $\mathcal{K}^{\perp}_{\epsilon,P}$ to $\mathcal{C}^{\perp}_{\epsilon,P}$. Suppose not. Then there exist $\phi \in K^{\perp}_{\epsilon,P}$, $\psi \in W^{2,t}(\Omega)$ such that

$$\Delta_y\phi - \phi + pA^{p-1}_{\epsilon,P}H^{-q}_{\epsilon,P}\phi + rc_{\epsilon,P}A^{r-1}_{\epsilon,P}H^{-s}_{\epsilon,P}\psi \in C^{\perp}_{\epsilon,P},$$

$$\Delta_x\psi - \psi - sc_{\epsilon,P}A^r_{\epsilon,P}H^{-s-1}_{\epsilon,P}\psi - qA^p_{\epsilon,P}H^{-q-1}_{\epsilon,P}\phi = 0,$$

$$\|\phi\|^2_{H^2(\Omega_{\epsilon,P})} + \|\psi\|^2_{W^{2,t}(\Omega)} = 1.$$

Similar to the proof of Proposition 4.2, we have

$$\psi(P) = -(1+o(1))c_{\epsilon,P}\frac{q}{s+1}\frac{\int_{\Omega_{\epsilon,P}}A^p_{\epsilon,P}\phi}{\int_{\Omega_{\epsilon,P}}A^r_{\epsilon,P}}$$

and substituting into the equation for ϕ we obtain

$$L_{\epsilon,P}\phi + o(1) \in C^{\perp}_{\epsilon,P}, \quad \phi \in K^{\perp}_{\epsilon,P}.$$

By Proposition 4.1, $\|\phi\|_{H^2(\Omega_{\epsilon,P})} = o(1)$, and hence $\|\psi\|_{W^{2,t}(\Omega)} = o(1)$. This is a contradiction! □

Now we are in a position to solve the equation

$$\pi_{\epsilon,P}\circ S_\epsilon\begin{pmatrix} A_{\epsilon,P}+\phi \\ H_{\epsilon,P}+\psi \end{pmatrix} = 0. \tag{4.17}$$

Since $\mathcal{L}_{\epsilon,P}|_{\mathcal{K}^{\perp}_{\epsilon,P}}$ is invertible (call the inverse $\mathcal{L}^{-1}_{\epsilon,P}$) we can rewrite

$$\Sigma = -(\mathcal{L}^{-1}_{\epsilon,P}\circ\pi_{\epsilon,P})\left(S_\epsilon\begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P}\end{pmatrix}\right) - (\mathcal{L}^{-1}_{\epsilon,P}\circ\pi_{\epsilon,P})N_{\epsilon,P}(\Sigma) \equiv M_{\epsilon,P}(\Sigma), \tag{4.18}$$

where

$$N_{\epsilon,P}(\Sigma) = S_\epsilon\begin{pmatrix} A_{\epsilon,P}+\phi \\ H_{\epsilon,P}+\psi\end{pmatrix} - S_\epsilon\begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P}\end{pmatrix} - S'_\epsilon\begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P}\end{pmatrix}\begin{bmatrix} \phi \\ \psi\end{bmatrix}$$

and the operator $M_{\epsilon,P}$ is defined by the last equation for $\Sigma \in H^2_N(\Omega_{\epsilon,P}) \times W^{2,t}(\Omega)$. We are going to show that the operator $M_{\epsilon,P}$ is a contraction on

$$B_{\epsilon,\delta} \equiv \{\Sigma \in H^2(\Omega_{\epsilon,P}) \times W^{2,t}(\Omega) | \|\Sigma\|_{H^2(\Omega_{\epsilon,P})\times W^{2,t}(\Omega)} < \delta\}$$

if δ is small enough. We have by Lemma 3.1 and Propositions 4.2 and 4.3,

$$\begin{aligned}
&\|M_{\epsilon,P}(\Sigma)\|_{H^2(\Omega_{\epsilon,P})\times W^{2,t}(\Omega)} \\
&\le \lambda^{-1}\left(\|\pi_{\epsilon,P}\circ N_{\epsilon,P}(\Sigma)\|_{L^2(\Omega_{\epsilon,P})\times L^t(\Omega)} + \left\|\pi_{\epsilon,P}\circ S_\epsilon\begin{pmatrix} A_{\epsilon,P} \\ H_{\epsilon,P}\end{pmatrix}\right\|_{L^2(\Omega_{\epsilon,P})\times L^t(\Omega)}\right) \\
&\le \lambda^{-1}C\left(c(\delta)\delta + \epsilon^{2t^{-1}-1}\frac{1}{\log\frac{1}{\epsilon}}\right) \qquad \text{(by Lemma 3.1)},
\end{aligned}$$

where $\lambda > 0$ is independent of $\delta > 0$ and $c(\delta) \to 0$ as $\delta \to 0$. Similarly we show

$$\|M_{\epsilon,P}(\Sigma) - M_{\epsilon,P}(\Sigma')\|_{H^2(\Omega_{\epsilon,P})\times W^{2,t}(\Omega)} \le \lambda^{-1}C(\epsilon^{1/2} + c(\delta)\delta)\|\Sigma - \Sigma'\|_{H^2(\Omega_{\epsilon,P})\times W^{2,t}(\Omega)},$$

where $c(\delta) \to 0$ as $\delta \to 0$. If we choose δ small enough, then $M_{\epsilon,P}$ is a contraction on $B_{\epsilon,\delta}$. The existence of a fixed point $\Sigma_{\epsilon,P}$ now follows from the contraction mapping principle and $\Sigma_{\epsilon,P}$ is a solution of (4.18).

We have thus proved Lemma 4.4.

LEMMA 4.4. *There exists $\bar\epsilon > 0$ such that for every pair of ϵ, P with $0 < \epsilon < \bar\epsilon$ there exists a unique $(\Phi_{\epsilon,P}, \Psi_{\epsilon,P}) \in \mathcal{K}^\perp_{\epsilon,P}$ satisfying*

$$S_\epsilon\left(\begin{pmatrix} A_{\epsilon,P} + \Phi_{\epsilon,P} \\ H_{\epsilon,P} + \Psi_{\epsilon,P}\end{pmatrix}\right) \in \mathcal{C}_{\epsilon,P}$$

and

$$\|(\Phi_{\epsilon,P}, \Psi_{\epsilon,P})\|_{H^2(\Omega_{\epsilon,P})\times W^{2,t}(\Omega)} \le C\epsilon^{2t^{-1}-1}. \tag{4.19}$$

We can improve the estimates in Lemma 4.4.

LEMMA 4.5. *Let $(\Phi_{\epsilon,P}, \psi_{\epsilon,P})$ be given by Lemma 4.4. Then we have*

$$\|\Phi_{\epsilon,P}\|_{L^\infty(\Omega_{\epsilon,P})} = O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right), \quad \|\Psi_{\epsilon,P}\|_{L^\infty(\Omega)} = O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right), \tag{4.20}$$

and

$$|\Psi_{\epsilon,P}(x) - \Psi_{\epsilon,P}(P)| \le C\frac{\epsilon}{(\log\frac{1}{\epsilon})^2}\log\left(\frac{|x-P|}{\epsilon} + 1\right) \quad \text{for } x \ne P. \tag{4.21}$$

Proof. The proof is divided into several steps.

First we note that by the equation for $\Phi_{\epsilon,P}$ and Lemmas 3.1 and 4.4,

$$\Delta_y\Phi_{\epsilon,P} - \Phi_{\epsilon,P} + pA^{p-1}_{\epsilon,P}H^{-q}_{\epsilon,P} - qA^p_{\epsilon,P}H^{-q-1}_{\epsilon,P}\Psi_{\epsilon,P} + f_1 \in C_{\epsilon,P},$$

where $\|f_1\|_{L^2(\Omega_{\epsilon,P})} = O(\frac{\epsilon}{\log\frac{1}{\epsilon}})$. Hence we obtain

$$\|\Phi_{\epsilon,P}\|_{H^2(\Omega_{\epsilon,P})} \le C\|A^p_{\epsilon,P}H^{-q-1}_{\epsilon,P}\Psi_{\epsilon,P}\|_{L^2(\Omega_{\epsilon,P})} + O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right)$$

$$\leq C\|\Psi_{\epsilon,P}(x)\|_{L^\infty(\Omega)} + O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right). \tag{4.22}$$

Next $\Psi_{\epsilon,P}$ satisfies

$$\Delta_x\Psi_{\epsilon,P} - \Psi_{\epsilon,P} = f_2 := c_{\epsilon,P}\frac{\hat{A}^r_{\epsilon,P}}{(\hat{H}_{\epsilon,P} - \psi_{\epsilon,P}(P))^s} - c_{\epsilon,P}\frac{(\hat{A}_{\epsilon,P} + \Phi_{\epsilon,P})^r}{(\hat{H}_{\epsilon,P} - \psi_{\epsilon,P}(x) + \Psi_{\epsilon,P})^s}.$$

We have

$$|f_2(x)| \leq Cc_{\epsilon,P}(w(y)^{r-1}|\Phi_{\epsilon,P}(y)| + w^r(y)|\Psi_{\epsilon,P}(x)|) + O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}c_{\epsilon,P}w^r(y)\right) \tag{4.23}$$

for a.e. $x \in \Omega$.

Therefore, we have by Lemma 1.2 and (4.22)

$$\Psi_{\epsilon,P}(x) = O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right) + O\left(\frac{1}{\log\frac{1}{\epsilon}}\|\Psi_{\epsilon,P}\|_{L^\infty(\Omega)}\right)$$

and so

$$\|\Psi_{\epsilon,P}\|_{L^\infty(\Omega)} = O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right)$$

or, equivalently,

$$\|\Psi_{\epsilon,P}\|_{L^\infty(\Omega_{\epsilon,P})} = O\left(\frac{\epsilon}{\log\frac{1}{\epsilon}}\right),$$

where $y = (x - P)/\epsilon$.

Moreover by Lemma 1.2 and (4.23),

$$\Psi_{\epsilon,P}(x) - \Psi_{\epsilon,P}(P) = O\left(\frac{\epsilon}{(\log\frac{1}{\epsilon})^2}|\log|y| + 1|\right).$$

Lemma 4.5 is proved. □

5. The reduced problem. In this section we solve the reduced problem and prove our main theorem.

By Lemma 4.4 there exists a unique solution $(\Phi_{\epsilon,P}, \psi_{\epsilon,P}) \in \mathcal{K}^\perp_{\epsilon,P}$ such that

$$S_\epsilon\begin{pmatrix} A_{\epsilon,P} + \Phi_{\epsilon,P} \\ H_{\epsilon,P} + \Psi_{\epsilon,P} \end{pmatrix} = \begin{pmatrix} v_{\epsilon,P} \\ 0 \end{pmatrix} \in \mathcal{C}_{\epsilon,P}.$$

Our idea is to find P such that

$$S_\epsilon\begin{pmatrix} A_{\epsilon,P} + \Phi_{\epsilon,P} \\ H_{\epsilon,P} + \Psi_{\epsilon,P} \end{pmatrix} \perp \mathcal{C}_{\epsilon,P}.$$

Let

$$W_{\epsilon,j}(P) := \frac{\log\frac{1}{\epsilon}}{\epsilon^2}\int_\Omega \left(S_1(A_{\epsilon,P} + \Phi_{\epsilon,P}, H_{\epsilon,P} + \Psi_{\epsilon,P})\frac{\partial A_{\epsilon,P}}{\partial P_j}\right),$$

$$W_\epsilon(P) := (W_{\epsilon,1}(P), \dots, W_{\epsilon,N}(P)).$$

Then $W_\epsilon(P)$ is a continuous map in P and our problem is reduced to finding a zero of the vector field $W_\epsilon(P)$.

Let us now calculate $W_\epsilon(P)$.

By Lemma 4.5,

$$\Psi_{\epsilon,P}(x) - \Psi_{\epsilon,P}(P) = O\left(\frac{\epsilon}{(\log\frac{1}{\epsilon})^2}\log\left(\frac{|x-P|}{\epsilon}+1\right)\right). \tag{5.1}$$

By (3.3) and (3.4), we have

$$\int_\Omega \left(S_1(A_{\epsilon,P}+\Phi_{\epsilon,P}, H_{\epsilon,P}+\psi_{\epsilon,P})\frac{\partial A_{\epsilon,P}}{\partial P_j}\right)$$

$$= \epsilon^2 \int_{\Omega_{\epsilon,P}} (\Delta_y\Phi_{\epsilon,P} - \Phi_{\epsilon,P} + pA_{\epsilon,P}^{p-1}H_{\epsilon,P}^{-q}\Phi_{\epsilon,P} - qA_{\epsilon,P}^{p-1}H_{\epsilon,P}^{-q-1}\Psi_{\epsilon,P})\frac{\partial A_{\epsilon,P}}{\partial P_j}$$

$$+O\left(\epsilon^3\left(\frac{1}{\log\frac{1}{\epsilon}}\right)^2\right)$$

$$+\epsilon^2\int_{\Omega_{\epsilon,P}} -q(\hat{A}_{\epsilon,P})^p(\hat{H}_{\epsilon,P})^{-q-1}[\psi_{\epsilon,P}(P+\epsilon y)-\psi_{\epsilon,P}(P)]\frac{\partial A_{\epsilon,P}}{\partial P_j}(y)\,dy$$

$$+O(e^{-d(P,\partial\Omega)/\epsilon}) = I_1 + I_2,$$

where I_1, I_2 are defined by the last equality.

For I_1, we note that $\|\Psi_{\epsilon,P}\|_{L^\infty(\Omega_{\epsilon,P}} = O(\frac{\epsilon}{\log\frac{1}{\epsilon}})$, $\frac{\partial A_{\epsilon,P}}{\partial P_j} = -\frac{1+o(1)}{\epsilon}\frac{\partial w}{\partial y_j}$, and hence

$$I_1 = \epsilon\int_{\Omega_{\epsilon,P}} (qA_{\epsilon,P}^{p-1}H_{\epsilon,P}^{-q-1}\Psi_{\epsilon,P})\frac{\partial w}{\partial y_j} + O\left(\epsilon^2\left(\frac{1}{\log\frac{1}{\epsilon}}\right)^2\right)$$

$$= \epsilon\int_{\Omega_{\epsilon,P}} (qw^{p-1}\Psi_{\epsilon,P})\frac{\partial w}{\partial y_j} + O\left(\epsilon^2\left(\frac{1}{\log\frac{1}{\epsilon}}\right)^2\right)$$

$$= \epsilon\int_{\Omega_{\epsilon,P}} (qw^{p-1}(y)H_{\epsilon,P}^{-q-1}(\Psi_{\epsilon,P}(P+\epsilon y)-\Psi_{\epsilon,P}(P)))\frac{\partial w}{\partial y_j} + O\left(\epsilon^2\left(\frac{1}{\log\frac{1}{\epsilon}}\right)^2\right)$$

$$= O\left(\epsilon^2\frac{1}{(\log\frac{1}{\epsilon})^2}\right)$$

by (5.1).

For I_2 we have

$$I_2 = C\epsilon\int_{\Omega_{\epsilon,P}} [\psi_{\epsilon,P}(P+\epsilon y)-\psi_{\epsilon,P}(P)]\frac{\partial w}{\partial y_j}\,dy\left(1+O\left(\frac{1}{\log\frac{1}{\epsilon}}\right)\right)$$

$$= C\frac{\epsilon}{\log\frac{1}{\epsilon}}\int_{R^2} -[H(P,P+\epsilon y)-H(P,P)]w'(|y|)\frac{y_i}{|y|}\,dy\left(1+O\left(\frac{1}{\log\frac{1}{\epsilon}}\right)\right)$$

$$= -C\frac{\epsilon^2}{\log\frac{1}{\epsilon}}\frac{\partial}{\partial P_j}H(P,P)\int_{R^2} w'(|y|)|y|\,dy + O\left(\frac{\epsilon^N}{(\log\frac{1}{\epsilon})^2}\right)$$

as $\epsilon \to 0$ uniformly in P, where $w'(|y|) = \frac{d}{dr}w(r)$ for $r = |y|$ and $C \neq 0$ denotes a generic constant.

Combining I_1 and I_2, we have

$$W_\epsilon(P) = c_0\nabla_P H(P,P) + O\left(\frac{1}{\log\frac{1}{\epsilon}}\right),$$

where $c_0 \neq 0$ is a generic constant.

Suppose at P_0 we have $\nabla_P H(P_0,P_0) = 0, \det(\nabla_j\nabla_k H(P_0,P_0)) \neq 0$; then the standard Brouwer fixed point theorem shows that for $\epsilon << 1$ there exists a P_ϵ such that $W_\epsilon(P_\epsilon) = 0$ and $P_\epsilon \to P_0$.

Thus we have proved the following proposition.

PROPOSITION 5.1. *For ϵ sufficiently small there exist points P_ϵ with $P_\epsilon \to P_0$ such that $W_\epsilon(P_\epsilon) = 0$.*

Finally, we prove Theorem 1.1.

Proof of Theorem 1.1. By Proposition 5.1, there exists $P_\epsilon \to P_0$ such that $W_\epsilon(P_\epsilon) = 0$. In other words, $S_1(A_{\epsilon,P_\epsilon} + \Phi_{\epsilon,P_\epsilon}, H_{\epsilon,P_\epsilon} + \Psi_{\epsilon,P_\epsilon}) = 0$. Let $\xi_\epsilon^{\frac{qr}{(p-1)(s+1)-qr}} = c_{\epsilon,P_\epsilon}, A_\epsilon = \xi_\epsilon^{q/(p-1)}(A_{\epsilon,P_\epsilon} + \Phi_{\epsilon,P_\epsilon}), H_\epsilon = \xi_\epsilon(H_{\epsilon,P} + \Psi_{\epsilon,P_\epsilon})$. It is easy to see that $H_\epsilon = 1 + O(\frac{1}{\log\frac{1}{\epsilon}}) > 0$ and hence $A_\epsilon \geq 0$. By the maximum principle, $A_\epsilon > 0$. Moreover A_ϵ, H_ϵ satisfy Theorem 1.1. □

Acknowledgment. Matthias Winter thanks the Department of Mathematics at The Chinese University of Hong Kong for their kind hospitality.

REFERENCES

[1] ADIMURTHI, G. MANCINNI, AND S.L. YADAVA, *The role of mean curvature in a semilinear Neumann problem involving the critical Sobolev exponent*, Comm. Partial Differential Equations, 20 (1995), pp. 591–631.

[2] ADIMURTHI, F. PACELLA, AND S.L. YADAVA, *Interaction between the geometry of the boundary and positive solutions of a semilinear Neumann problem with critical nonlinearity*, J. Funct. Anal., 118 (1993), pp. 318–350.

[3] ADIMURTHI, F. PACELLA, AND S.L. YADAVA, *Characterization of concentration points and L^∞-estimates for solutions involving the critical Sobolev exponent*, Differential Integral Equations, 8 (1995), pp. 41–68.

[4] P. BATES AND G. FUSCO, *Equilibria with Many Nuclei for the Cahn–Hilliard Equation*, preprint, 1997.

[5] P. BATES, E.N. DANCER, AND J. SHI, *Multi-spike stationary solutions of the Cahn–Hilliard equation in higher-dimension and instability*, Adv. Differential Equations, 4 (1999), pp. 1–69.

[6] G. CERAMI AND J. WEI, *Multiplicity of multiple interior spike solutions for some singularly perturbed Neumann problem*, International Math. Res. Notices, 12 (1998), pp. 601–626.

[7] X. CHEN AND M. KOWALCZYK, *Slow Dynamics of Interior Spikes in the Shadow Gierer–Meinhardt System*, preprint, 1999.

[8] M. DEL PINO, *A priori estimates and applications to existence-nonexistence for a semilinear elliptic system*, Indiana Univ. Math. J., 43 (1994), pp. 703–728.

[9] M. DEL PINO, P. FELMER, AND M. KOWALCZYK, *Boundary Spikes in the Gierer–Meinhardt System*, preprint, 1999.

[10] M. DEL PINO, P. FELMER, AND J. WEI, *On the role of mean curvature in some singularly perturbed Neumann problems*, SIAM J. Math. Anal., to appear.

[11] A. FLOER AND A. WEINSTEIN, *Nonspreading wave packets for the cubic Schrödinger equation with a bounded potential*, J. Funct. Anal., 69 (1986), pp. 397–408.

[12] C. Gui and N. Ghoussoub, *Multi-peak solutions for a semilinear Neumann problem involving the critical Sobolev exponent*, Math. Z., 229 (1998), pp. 443–474.
[13] C. Gui and N. Ghoussoub, *New variational principles and multi-peak solutions for Neumann problems involving the critical Sobolev exponent*, in Canadian Mathematical Society, 1945–1995, vol. 3, J.B. Carrell and R. Murty, eds., Canadian Mathematical Society, Ottawa, ON, 1996, pp. 125–152.
[14] A. Gierer and H. Meinhardt, *A theory of biological pattern formation*, Kybernetik (Berlin), 12 (1972), pp. 30–39.
[15] C. Gui, *Multi-peak solutions for a semilinear Neumann problem*, Duke Math. J., 84 (1996), pp. 739–769.
[16] B. Gidas, W.M. Ni, and L. Nirenberg, *Symmetry of positive solutions of nonlinear elliptic equations in R^N*, Adv. Math. Suppl Stud., 7A (1981), pp. 369–402.
[17] C. Gui, J. Wei, and M. Winter, *Multiple boundary peak solutions for some singularly perturbed Neumann problems*, Ann. Inst. H. Poincaré Anal. Non Linéaire, to appear.
[18] C. Gui and J. Wei, *Multiple interior peak solutions for some singular perturbation problems*, J. Differential Equations, to appear.
[19] D. Iron and M. Ward, *A metastable spike solution for a non-local reaction-diffusion model*, SIAM J. Appl. Math., submitted.
[20] K.F. Keller and L.A. Segal, *Initiation of slime mold aggregation viewed as an instability*, J. Theoret. Biol., 26 (1970), pp. 399–415.
[21] M. Kowalczyk, *Multiple spike layers in the shadow Gierer–Meinhardt system: Existence of equilibria and the quasi-invariant manifold*, Duke Math. J., 98 (1999), pp. 59–111.
[22] Y.-Y. Li, *On a singularly perturbed equation with Neumann boundary condition*, Comm. Partial Differential Equations, 23 (1998), pp. 487–545.
[23] J.P. Keener, *Activators and inhibitors in pattern formation*, Stud. Appl. Math., 59 (1978), pp. 1–23.
[24] M.K. Kwong and L. Zhang, *Uniqueness of positive solutions of $\Delta u + f(u) = 0$ in an annulus*, Differential Integral Equations, 4 (1991), pp. 583–599.
[25] C.-S. Lin and W.-M. Ni, *On the diffusion coefficient of a semilinear Neumann problem*, Lecture Notes in Math. 1340, Springer, New York, 1988, pp. 160–174.
[26] W.-M. Ni, *Diffusion, cross-diffusion, and their spike-layer steady states*, Notices Amer. Math. Soc., 45 (1998), pp. 9–18.
[27] W.-M. Ni, X. Pan, and I. Takagi, *Singular behavior of least-energy solutions of a semilinear Neumann problem involving critical Sobolev exponents*, Duke Math. J., 67 (1992), pp. 1–20.
[28] W.-M. Ni and I. Takagi, *On the Neumann problem for some semilinear elliptic equations and systems of activator-inhibitor type*, Trans. Amer. Math. Soc., 297 (1986), pp. 351–368.
[29] W.-M. Ni and I. Takagi, *On the shape of least energy solution to a semilinear Neumann problem*, Comm. Pure Appl. Math., 41 (1991), pp. 819–851.
[30] W.-M. Ni and I. Takagi, *Locating the peaks of least energy solutions to a semilinear Neumann problem*, Duke Math. J., 70 (1993), pp. 247–281.
[31] W.-M. Ni and I. Takagi, *Point-condensation generated by a reaction-diffusion system in axially symmetric domains*, Japan J. Indust. Appl. Math., 12 (1995), pp. 327–365.
[32] W.-M. Ni, I. Takagi, and E. Yanagida, in preparation.
[33] W.-M. Ni and J. Wei, *On the location and profile of spike-layer solutions to singularly perturbed semilinear Dirichlet problems*, Comm. Pure. Appl. Math., 48 (1995), pp. 731–768.
[34] Y. Nishiura, *Global structure of bifurcating solutions of some reaction-diffusion systems*, SIAM J. Math. Anal., 13 (1982), pp. 555–593.
[35] Y.G. Oh, *Existence of semi-classical bound states of nonlinear Schrödinger equations with potentials of the class $(V)_a$*, Comm. Partial Differential Equations, 13 (1988), pp. 1499–1519.
[36] Y.G. Oh, *On positive multi-bump bound states of nonlinear Schrödinger equations under multiple-well potentials*, Comm. Math. Phys., 131 (1990), pp. 223–253.
[37] I. Takagi, *Point-condensation for a reaction-diffusion system*, J. Differential Equations, 61 (1986), pp. 208–249.
[38] M.J. Ward, *An asymptotic analysis of localized solutions for some reaction-diffusion models in multi-dimensional domains*, Stud. Appl. Math., 97 (1996), pp. 103–126.
[39] J. Wei, *On the boundary spike layer solutions of singularly perturbed semilinear Neumann problem*, J. Differential Equations, 134 (1997), pp. 104–133.
[40] J. Wei, *On the interior spike layer solutions of singularly perturbed semilinear Neumann problem*, Tohoku Math. J., 50 (1998), pp. 159–178.
[41] J. Wei, *On the interior spike layer solutions for some singular perturbation problems*, Proc. Roy. Soc. Edinburgh Sect. A, 128 (1998), pp. 849–874.
[42] J. Wei, *Uniqueness and Eigenvalue Estimates of Boundary Spike Solutions*, preprint, 1998.
[43] J. Wei, *On single interior spike solutions of Gierer–Meinhardt system: Uniqueness and spectrum*

estimates, European J. Appl. Math., to appear.
[44] J. Wei and M. Winter, *Stationary solutions for the Cahn–Hilliard equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 15 (1998), pp. 459–492.
[45] J. Wei and M. Winter, *On the Cahn–Hilliard equations* II*: Interior spike layer solutions*, J. Differential Equations, 148 (1998), pp. 231–267.
[46] J. Wei and M. Winter, *Multiple boundary spike solutions for a wide class of singular perturbation problems*, J. London Math. Soc., to appear.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1264–1281

EXISTENCE OF THE GLOBAL CLASSICAL SOLUTION FOR A TWO-PHASE STEFAN PROBLEM*

M. A. BORODIN†

Abstract. In this work we prove the existence of the global classical solution in a two-phase multidimensional Stefan problem. We apply a method which consists of the following. First, we construct a special system of difference–differential approximating elliptic problems, then we prove some uniform estimates and pass to the limit. We prove that the free boundary is given by the graph of a function from the $H^{2+\alpha,1+\frac{\alpha}{2}}$ class.

Key words. two-phase Stefan problem, global classic solution, free boundary problem

AMS subject classification. 35R35

PII. S0036141097332530

Introduction. In this work we study the existence of the global classical solution for a two-phase multidimensional Stefan problem; see, e.g., [21, chap. 5, §9]. Let $D = \{x \in \mathbf{R^3} : \quad 0 < \mathrm{R}_1 < |\mathrm{x}| < \mathrm{R}_2\}$, $D_T = D \times (0,T)$, $B_i = \{x \in \mathbf{R^3} : |\mathrm{x}| < \mathrm{R_i}\}$; $\mathrm{i} = 1,2$, $T > 0$ is a fixed number. The problem is to find a function $u(x,t)$ and domains Ω_T, G_T, which satisfy

$$\Delta u - \frac{\partial u}{\partial t} = 0 \quad \text{in} \quad \Omega_T \cup G_T, \tag{0.1}$$

$$\Omega_T = \{(x,t) \in D_T : 0 < u(x,t) < 1\}, \quad G_T = \{(x,t) \in D_T : u(x,t) > 1\}.$$

On the known boundary

$$u(x,t) = \varphi_i(x,t) \quad \text{on} \quad \partial B_i \times (0,T),\ i = 1,2. \tag{0.2}$$

On the unknown (free) boundary $\gamma_T = \partial\Omega_T \cap D_T = \partial G_T \cap D_T$

$$u^+ = u^- = 1, \quad \sum_{k=1}^{3} \left(\frac{\partial u^-}{\partial x_i} - \frac{\partial u^+}{\partial x_i} \right) \cos(n, x_i) + \lambda \cos(n,t) = 0, \tag{0.3}$$

where λ is a positive constant, n is the normal to the surface γ_T directed to the side of increase of the function $u(x,t)$, and $u^+(x,t)$, $u^-(x,t)$ are the boundary values on the surface γ_T taken from the domains G_T, Ω_T, respectively.

The initial conditions are

$$u(x,0) = \psi(x) \quad \text{in} \quad \overline{D}, \ \ \psi(x) = \varphi_i(x,0) \quad \text{on} \quad \partial B_i, \tag{0.4}$$

$$0 \le \psi(x) < 1 \quad \text{on} \quad \partial B_1, \quad \psi(x) > 1 \quad \text{on} \quad \partial B_2,$$

$$\Omega_0 = \{x \in D : \psi(x) < 1\}, \quad G_0 = \{x \in D : \psi(x) > 1\}, \ \gamma_0 = \partial\Omega_0 \cap D = \partial G_0 \cap D.$$

*Received by the editors January 15, 1998; accepted for publication (in revised form) February 4, 1999; published electronically October 4, 1999.
http://www.siam.org/journals/sima/30-6/33253.html
†Department of Mathematics, Donetsk State University, University Street 24, Donetsk, Ukraine (borodin@univ.donetsk.ua).

Problem (0.1)–(0.4) represents a mathematical model that describes the spreading of heat in a medium with a varying phase state. The function $u(x,t)$ is interpreted as the temperature of the medium, γ_T is the interface between the liquid and solid phases, and $u(x,t)=1$ is the temperature of melting.

Multidimensional Stefan problems have been studied by many authors. The concept of weak solutions was introduced in [1], [2], where the existence and uniqueness of such solutions have been proved. Certain free boundary problems have been reduced to variational inequalities [3], [4]. Such reduction allowed us to prove [5] Lipschitz continuity of the free boundary in a one-phase Stefan problem. However, this was not enough to prove the existence of a classical solution. The existence of a classical solution in a one-phase Stefan problem was proved in [6]. The works by L. Caffarelli [7]–[9] played an important role here. A variational inequality equivalent to a two-phase Stefan problem was obtained by M. Fremon [10], but this allowed us to prove just continuity of the temperature.

Lipschitz continuity of the free boundary in a two-phase problem was proved in [11]. Moreover, in [12] it was shown that a viscosity solution with Lipschitz free boundary under certain nondegeneracy conditions is, actually, classical, and the free boundary is a C^1 graph in space and time. The existence of classical solutions for a small time interval was proved in [13]–[15]. The main result of this paper is the following theorem.

THEOREM 0.1. *Let the following conditions be satisfied:*

$$\psi(x)\in C^{2+\alpha}(\overline{D}),\quad \Delta\psi\le 0\quad in\quad \overline{D},\quad \frac{\partial\psi}{\partial\rho}>0\quad in\quad \overline{D},$$

$$\varphi_1(x,t)=0\quad on\quad \partial B_1,\quad \varphi_2(x,t)=q=\mathrm{const}>1\quad on\quad \partial B_2,$$

and assume that the corresponding compatibility conditions at $t=0$, $x\in\partial\Omega_T\cup\partial G_T$ hold. Then $\forall\, T>0$ there exists a unique solution of the problem (0.1)–(0.4) *and*

$$u(x,t)\in C(\overline{D}_T)\cap\Big(H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{\Omega}_T)\times H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{G}_T)\Big);$$

the free boundary is given by the graph $\rho=\omega(\theta_1,\theta_2,t)$ of a function $\omega(\theta_1,\theta_2,t)\in H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{\Pi}_T)$, where (ρ,θ_1,θ_2) are spherical coordinates, $\Pi=\{(\theta_1,\theta_2):0<\theta_1<2\pi, 0<\theta_2<\pi\}$, $\Pi_T=\Pi\times(0,T)$.

This result has been announced in [16]. The theorem provides sufficient conditions for the existence of a global classical solution for a Stefan problem. Note that similar assumptions on the data have been used in [11]. Although our conditions are not necessary, examples [17], [18], [11], [12] show that the smoothness of initial and boundary conditions alone does not imply the smoothness of the free boundary.

The paper is organized as follows. In sections 1 and 2 we construct a difference–differential approximation of our problem and study its properties. Uniform estimates for the approximate solutions are obtained in section 3. In section 4 we pass to the limit and prove the main theorem.

In [19] we have also considered a contact Stefan problem with Neumann boundary conditions on the fixed part of the boundary.

For the sake of simplicity we consider three-dimensional space; the same method can be applied in the case of arbitrary space dimension.

1. Construction of the approximating problems. Let us construct a system of approximating problems. We divide the cylinder D_T by the planes $t=kh$, $1\le$

$k \le N$; N is a positive number. For any number $\varepsilon > 0$ we introduce a function $\chi_\varepsilon \in C^\infty(\mathbf{R^1})$ such that

$$\chi_\varepsilon(x) = 1 \quad \forall x \le 1, \quad \chi_\varepsilon(x) = 0 \quad \forall x \ge 1 + \varepsilon, \quad \chi'_\varepsilon(x) \le 0.$$

Define the functions $\{u_k(x,h,\varepsilon)\}$, $\{F_k(x,h,\varepsilon)\}$ as solutions of the following problem.

$$\Delta u_k - \frac{u_k - u_{k-1}}{h} = -\lambda \frac{\chi_\varepsilon(u_k) - \chi_\varepsilon(u_0)}{h} + \frac{F_{k-1}}{h} \text{ in } D,\ k = 1, 2, \dots, N, \tag{1.1}$$

$$u_k = \varphi_i(x, kh) \text{ on } \partial B_i,\ u_0 = \psi(x) \text{ in } D, \tag{1.2}$$

$$\Delta F_k - \frac{F_k}{h} = -\lambda \frac{\chi_\varepsilon(u_k) - \chi_\varepsilon(u_0)}{h} \text{ in } D, \tag{1.3}$$

$$F_k = 0 \text{ on } \partial B_1 \cup \partial B_2,\ F_0 = 0 \text{ in } D. \tag{1.4}$$

If $h > 0, \varepsilon > 0$ are fixed, the solvability of problem (1.1)–(1.4) is evident. It can be considered step by step starting with $k = 1$. First, we find the function $F_{k-1}(x,h,\varepsilon)$ ($F_0 = 0$), then we put it to the right-hand side of (1.1) and consider the boundary problem for the function $u_k(x,h,\varepsilon)$. We plug the function thus obtained into the right-hand side of (1.3) and find the function $F_k(x,h,\varepsilon)$, and so on. The existence of a solution for each of the above-mentioned problems is known [20, chap. 3, §1]. Thus, we obtain the following statement.

THEOREM 1.1. *Let the following conditions hold:*

$$\psi(x) \in C^{2+\alpha}(\overline{D}), \quad \varphi_i(x, kh) \in C^{2+\alpha}(\overline{D}), \quad \alpha \in (0,1),$$

and the functions $\psi(x)$, $\varphi_i(x,kh)$ on $x \in \partial B_i$, and $k = 0$ satisfy the corresponding compatibility conditions. Then problem (1.5)–(1.8) *is solvable and*

$$u_k(x,h,\varepsilon) \in C^{2+\alpha}(\overline{D}), \quad F_k(x,h,\varepsilon) \in C^{2+\alpha}(\overline{D}).$$

In what follows we shall show that the linear interpolations of the functions $\{u_k(x,h,\varepsilon)\}$ with respect to t converge to a solution of the Stefan problem (0.1)–(0.4) as $\varepsilon \to 0$, $h \to 0$.

Subtract (1.3) from (1.1) and set

$$w_k(x,h,\varepsilon) = u_k(x,h,\varepsilon) - F_k(x,h,\varepsilon) \tag{1.5}$$

Then the functions $\{w_k(x,h,\varepsilon)\}$ will satisfy the following problem.

$$\Delta w_k - \frac{w_k - w_{k-1}}{h} = 0 \quad \text{in} \quad D, \tag{1.6}$$

$$w_k(x,h,\varepsilon) = \varphi_i(x,kh) \quad \text{on} \quad \partial B_i, \quad w_0 = \psi(x) \quad \text{in} \quad D. \tag{1.7}$$

The equality (1.5) shows that the functions $u_k(x,h,\varepsilon)$ are sums of two summands; the first one $w_k(x,h,\varepsilon)$ is more smooth than the solution of the Stefan problem, while the second one $F_k(x,h,\varepsilon)$ contains information about the behavior of the solution near the free boundary. However, this second summand is the solution of a sufficiently simple problem. The exact meaning of these words will be clear from what follows.

2. Properties of $\{w_k(x,h,\varepsilon)\}$.

THEOREM 2.1. *Let the assumptions of Theorem* 1.1 *hold and suppose that*

$$|\varphi_i(x,(k-1)h)-\varphi_i(x,kh)|\le ch \quad \forall x\in\partial B_i, \tag{2.1}$$

where the constant c does not depend on h and k. Then there exists a constant M, such that

$$|w_{k-1}(x,h,\varepsilon)-w_k(x,h,\varepsilon)|\le Mh \quad \forall x\in\overline{D}, \tag{2.2}$$

where M does not depend on k, h, and ε.
If instead of (2.1) *the following condition holds:*

$$\Delta\psi<0 \quad in \quad \overline{D},\ 0<c_1h\le\varphi(x,(k-1)h)-\varphi(x,kh)\le c_2h, \tag{2.1*}$$

then

$$0<M_1h\le w_{k-1}(x,h,\varepsilon)-w_k(x,h,\varepsilon)\le M_2h, \tag{2.2*}$$

where the constants c_i, M_i do not depend on k, h, and ε.

Proof. On the boundary of the domain D the estimation (2.1) follows from the hypothesis of the theorem. Let us assume that

$$\max_{1\le k\le N,x\in\overline{D}}[w_{k-1}(x,h,\varepsilon)-w_k(x,h,\varepsilon)]=w_{n-1}(x_0,h,\varepsilon)-w_n(x_0,h,\varepsilon),$$

where $x_0\in D$. At a local maximum point we have $\Delta(w_{n-1}-w_n)\le 0$. Therefore, the equation

$$\Delta(w_{n-1}-w_n)-\frac{w_{n-1}-w_n}{h}=-\frac{w_{n-2}-w_{n-1}}{h} \tag{2.2}$$

implies

$$w_{n-1}(x_0,h,\varepsilon)-w_n(x_0,h,\varepsilon)\le w_{n-2}(x_0,h,\varepsilon)-w_{n-1}(x_0,h,\varepsilon).$$

From this estimate we conclude that

$$w_{n-1}(x_0,h,\varepsilon)-w_n(x_0,h,\varepsilon)\le w_0(x_0,h,\varepsilon)-w_1(x_0,h,\varepsilon),$$

but the function $w_0(x,h,\varepsilon)-w_1(x,h,\varepsilon)$ satisfies the equation

$$\Delta(w_0-w_1)-\frac{w_0-w_1}{h}=\Delta\psi. \tag{2.3}$$

Therefore,

$$\max_{1\le k\le N,x\in\overline{D}}[w_{k-1}(x,h,\varepsilon)-w_k(x,h,\varepsilon)]\le\max_{x\in\overline{D}}|\Delta\psi|h.$$

Similarly, we estimate the minimum of the functions $\{w_{k-1}(x,h,\varepsilon)-w_k(x,h,\varepsilon)\}$. The second part of the theorem is obvious. □

COROLLARY 2.1. *Let the assumptions of Theorem* 2.1 *hold and*

$$\|\varphi_i(x,kh)\|_{C^{2+\alpha}(\overline{D})}\le c_3.$$

Then $\exists M_3 > 0$, *such that*

$$\|w_k(x,h,\varepsilon)\|_{C^{1+\alpha}(\overline{D})} \le M_3, \tag{2.4}$$

where the constant M_3 *does not depend on* h, ε, *and* k.

Proof. From (2.1) and (1.6) we get

$$0 < M_1 \le -\Delta w_k \le M_2.$$

After that the estimation (2.4) follows from known properties of solutions of elliptic boundary problems. □

THEOREM 2.2. *Let the assumptions of Theorem* 1.1 *hold and*

$$\|\varphi_i(x,kh)\|_{C^{2+\alpha}(\overline{D})} + \Big\|\frac{\varphi_i[x,(k-1)h] - \varphi_i(x,kh)}{h}\Big\|_{C^{\alpha}(\overline{D})} \le c_4.$$

Then $\exists\ M_4 > 0$, *such that*

$$\|w_k\|_{C^{2+\alpha}(\overline{D})} + \Big\|\frac{w_{k-1}(x,h,\varepsilon) - w_k(x,h,\varepsilon)}{h}\Big\|_{C^{\alpha}(\overline{D})} \le M_4, \tag{2.5}$$

where the constant M_4 *does not depend on* h, ε, *and* k.

Proof. We use the method suggested in [20]. Let $\zeta_1(x), \zeta_2(x), \dots, \zeta_l(x)$ be nonnegative indefinitely differentiable functions with compact supports such that their sum is identically equal to one on $\overline{D}$, i.e.,

$$\sum_{k=1}^{l} \zeta_k(x) = 1, \quad x \in \overline{D}.$$

The functions $\{w_k(x,h,\varepsilon) - w_{k-1}(x,h,\varepsilon)\}$ have the form

$$\sum_{s=1}^{l} [w_k^s(x,h,\varepsilon) - w_{k-1}^s(x,h,\varepsilon)] = w_k(x,h,\varepsilon) - w_{k-1}(x,h,\varepsilon),\ k = 1,2,\dots,N,$$

where

$$w_k^s(x,h,\varepsilon) - w_{k-1}^s(x,h,\varepsilon) = \zeta_s(x)[w_k(x,h,\varepsilon) - w_{k-1}(x,h,\varepsilon)].$$

If the support of $\zeta_s(x)$ lies in the domain D, we can consider $\{w_k^s(x,h,\varepsilon) - w_{k-1}^s(x,h,\varepsilon)\}$ as compactly supported functions from $C^{2+\alpha}(\mathbf{R^3})$, satisfying the equations

$$\Delta(w_k^s - w_{k-1}^s) - \frac{w_k^s - w_{k-1}^s}{h} = -\frac{w_{k-1}^s - w_{k-2}^s}{h} - (f_k^s - f_{k-1}^s), \tag{2.6}$$

where $f_k^s = -2\nabla w_k \nabla \zeta_s - w_k \Delta \zeta_s$. Let x_0 belong to the support of the function $\zeta_s(x)$ and $K_R(x_0)$ be the ball with its center at the point x_0 and the radius R so large that its boundary lies outside of the support of $\zeta_s(x)$.

Let

$$\Gamma_{n-k+1}(|x-y|) = \frac{ih}{2\pi} \oint\limits_{\partial L^+(\rho)} \frac{Sh[\sqrt{z}(R-|x-y|)]}{4\pi|x-y|Sh(\sqrt{z}R)} \frac{dz}{(1-zh)^{n-k+1}}, \tag{2.7}$$

where $L^+(\rho) = \{z = \xi + i\eta : Rez > -\frac{\pi^2}{R^2}, |z| < \rho\}$, $\partial L^+(\rho)$ is the boundary of this set, $\rho \geq \frac{2}{h}, Sh(x) = \frac{e^x - e^{-x}}{2}$. The numerator and the denominator of the integrand have the same branch point. Therefore, in the domain $L^+(\rho)$ it is possible to choose a univalent branch of the integrand by setting, for example, $\sqrt{1} = 1$. We shall multiply both sides of (2.6) by $\Gamma_{n-k+1}(|x_0 - y|)$, sum up over k, $1 \leq k \leq n$, and then integrate over $K_R(x_0) \setminus K_\delta(x_0)$. After that we shall make the passage to the limit $\delta \to 0$. Thus, we shall construct an integral representation of a solution of (2.6)

$$w_n^s(x_0, h, \varepsilon) - w_{n-1}^s(x_0, h, \varepsilon) = \int\limits_{K_R(x_0)} \zeta_s(y) \Delta\psi \Gamma_n(|x_0 - y|)\, dy + \sum_{k=1}^{n} \int\limits_{K_R(x_0)} (f_k^s - f_{k-1}^s) \Gamma_{n-k+1}(|x_0 - y|)\, dy.$$

Let us transform the last term

$$\sum_{k=1}^{n} \int\limits_{K_R(x_0)} (f_k^s - f_{k-1}^s) \Gamma_{n-k+1}(|x_0 - y|)\, dy = - \int\limits_{K_R(x_0)} f_0^s \Gamma_n(|x_0 - y|)\, dy$$
$$+ \sum_{k=1}^{n} \int\limits_{K_R(x_0)} f_k^s [\Gamma_{n-k+1}(|x_0 - y|) - \Gamma_{n-k}(|x_0 - y|)]\, dy, \quad \Gamma_0(|x_0 - y|) \equiv 0.$$

Thus, we obtain the following integral representation:

$$w_n^s(x_0, h, \varepsilon) - w_{n-1}^s(x_0, h, \varepsilon) = \int\limits_{K_R(x_0)} (\zeta_s \Delta\psi - f_0^s) \Gamma_n(|x_0 - y|)\, dy + \sum_{k=1}^{n} \int\limits_{K_R(x_0)} f_k^s [\Gamma_{n-k+1}(|x_0 - y| - \Gamma_{n-k}(|x_0 - y|)]\, dy.$$

Let y_0 be a point in the support of the function $\zeta_s(x)$ such that $x_0 \neq y_0$ and $v_n^s(x, h, \varepsilon) = w_n^s(x, h, \varepsilon) - w_{n-1}^s(x, h, \varepsilon)$. Then for the difference

$$v_n^s(x_0, h, \varepsilon) - v_n^s(y_0, h, \varepsilon)$$

after several simplifications we get the following expression:

(2.8)

$$v_n^s(x_0, h, \varepsilon) - v_n^s(y_0, h, \varepsilon) = \int\limits_{K_R(x_0)} [\phi_s(y, h, \varepsilon) - \phi_s(y - x_0 + y_0, h, \varepsilon)] \Gamma_n(|x_0 - y|)\, dy$$
$$+ \sum_{k=1}^{n} \int\limits_{K_R(x_0)} [f_k^s(y, h, \varepsilon) - f_k^s(y - x_0 + y_0, h, \varepsilon)][\Gamma_{n-k+1}(|x_0 - y|) - \Gamma_{n-k}(|x_0 - y|)]\, dy,$$

where $\phi_s = \zeta_s \Delta\psi - f_0^s$. Note the following properties of the fundamental solutions $\{\Gamma_n(|x_0 - y|)\}$. Let $y \neq x_0$, then

$$\Delta\Gamma_1 - \frac{\Gamma_1}{h} = 0, \quad \Gamma_1(|x_0 - y|) = \frac{Sh\left[\sqrt{\frac{1}{h}}(R - |x_0 - y|)\right]}{4\pi |x_0 - y| Sh\left(\sqrt{\frac{1}{h}} R\right)},$$
$$\Delta\Gamma_{n-k+1} - \frac{\Gamma_{n-k+1} - \Gamma_{n-k}}{h} = 0, \quad \Gamma_{n-k+1}(R) = 0, \quad k = 1, 2, \ldots, n-1.$$

These formulas can be easily proved. They imply that

$$\Gamma_{n-k+1}(|x_0-y|)>0,$$

$$\int\limits_{K_R(x_0)} \frac{\Gamma_{n-k+1}(|x_0-y|)-\Gamma_{n-k}(|x_0-y|)}{h}\,dy = \int\limits_{\partial K_R(x_0)} \frac{\partial\Gamma_{n-k+1}}{\partial\nu}\,ds<0,$$

where ν is the outward pointing field to $\partial K_R(x_0)$. Therefore,

$$\int\limits_{K_R(x_0)} \frac{\Gamma_n(|x_0-y|)}{h}\,dy \le \int\limits_{K_R(x_0)} \frac{\Gamma_1(|x_0-y|)dy}{h} = 1-\frac{\sqrt{\frac{1}{h}}R}{Sh\sqrt{\frac{1}{h}}R}\le 1. \tag{2.9}$$

Let

$$\Gamma^1_{n-k+1}(|x_0-y|)=\frac{ih^2}{2\pi}\oint\limits_{\partial L^+(\rho)} \frac{zSh[\sqrt{z}(R-|x_0-y|)]}{4\pi|x_0-y|Sh(\sqrt{z}R)}\frac{dz}{(1-zh)^{n-k+1}},$$

then it follows from (2.7) that

$$\Gamma_{n-k+1}(|x_0-y|)-\Gamma_{n-k}(|x_0-y|)=\Gamma^1_{n-k+1}(|x_0-y|). \tag{2.10}$$

After that it is obvious that

$$\sum_{k=1}^{n}\int\limits_{K_R(x_0)} \frac{|\Gamma^1_{n-k+1}(|x_0-y|)|}{h}\,dy\le c, \tag{2.11}$$

where the constant c does not depend on h. From (2.8), taking into account the estimates (2.9) and (2.11), we obtain

$$\begin{aligned}|v^s_n(x_0,h,\varepsilon)-&v^s_n(y_0,h,\varepsilon)|\le c_1|x_0-y_0|^\alpha \int\limits_{K_R(x_0)}\Gamma_n(|x_0-y|)\,dy\\ &+\int\limits_{K_R(x_0)}\max_{1\le k\le n}|f^s_k(y,h,\varepsilon)-f^s_k(y-x_0+y_0,h,\varepsilon)|\sum_{k=1}^{n}\left|\Gamma^1_{n-k+1}(|x_0-y|)\right|dy\\ &\le c_1h|x_0-y_0|^\alpha+c_2h|x_0-y_0|^\alpha,\end{aligned}$$

where the constants c_1, c_2 do not depend on h, ε, n. Let the support of the function $\zeta_s(x)$ lie only partially in the domain D. By a parallel transport we may force the part of ∂D which belongs to the support to pass through the origin. The inversion $x\to y=x\setminus|x|^2$ is a diffeomorphism of $\mathbf{R^3}\setminus\{0\}$ onto itself and it maps a ball, such that its boundary passes through the origin onto a half-space. Let us denote by $\overline{v}^s_k$ the Kelvin transformation

$$\overline{v}^s_k(x,h,\varepsilon)=|x|^{-1}v^s_k(x/|x|^2,h,\varepsilon).$$

We take the fundamental solutions $\{\Gamma_{n-k+1}(|x_0-y|)-\Gamma_{n-k}(|x^*_0-y|)\}$, where x^*_0 is symmetric to x_0 with respect to the flat part of the boundary, and for $\overline{v}^s_k$ we repeat with small modifications the previous reasoning. After that, for the functions $\{w_k(x,h,\varepsilon)\}$ we get

$$\Delta w_k(x,h,\varepsilon)=\frac{w_k(x,h,\varepsilon)-w_{k-1}(x,h,\varepsilon)}{h}\in C^\alpha(\overline{D}).$$

Therefore, taking into account known properties of solutions of elliptic problems, we obtain the estimation (2.5). □

3. Uniform estimations of $\{u_k(x,h,\varepsilon)\}$.

THEOREM 3.1. *Let the assumptions of Theorem* 1.1 *and the estimation* (2.1) *hold. Then*

$$|u_k(x,h,\varepsilon) - u_{k-1}(x,h,\varepsilon)| \le M_1 h. \tag{3.1}$$

If instead of (2.1) *the condition* (2.1*) *holds, then*

$$0 \le u_{k-1}(x,h,\varepsilon) - u_k(x,h,\varepsilon) \le M_2 h, \tag{3.2*}$$

where the constants M_i do not depend on k, h, and ε.

Proof. In view of (2.1), it suffices to prove the estimation (3.1) in D. Using (1.5), we present (1.1) as follows:

$$\Delta u_k - \frac{u_k}{h} = -\frac{w_{k-1}}{h} - \lambda \frac{\chi_\varepsilon(u_k) - \chi_\varepsilon(u_0)}{h}. \tag{3.2}$$

Let us write down the equation for the difference $u_{k-1}(x,h,\varepsilon) - u_k(x,h,\varepsilon)$:

$$\Delta(u_{k-1} - u_k) - \frac{u_{k-1} - u_k}{h} = -\frac{w_{k-2} - w_{k-1}}{h} + \lambda \frac{\chi_\varepsilon(u_k) - \chi_\varepsilon(u_{k-1})}{h}. \tag{3.3}$$

Let us assume that the function $u_{k-1}(x,h,\varepsilon) - u_k(x,h,\varepsilon)$ has a negative minimum and attains it at an interior point. Then at this point $\Delta(u_{k-1} - u_k) \ge 0$. Therefore, (3.3) implies

$$u_{k-1} - u_k \ge w_{k-2} - w_{k-1} - \lambda[\chi_\varepsilon(u_k) - \chi_\varepsilon(u_{k-1})] \ge w_{k-2} - w_{k-1}.$$

If the function $u_{k-1}(x,h,\varepsilon) - u_k(x,h,\varepsilon)$ has a local maximum at an interior point, then at this point

$$u_{k-1} - u_k \le w_{k-2} - w_{k-1} - \lambda[\chi_\varepsilon(u_k) - \chi_\varepsilon(u_{k-1})] \le w_{k-2} - w_{k-1}.$$

The proof of the second part of the theorem is quite similar. □

Let us estimate $|\nabla u_k|$ in $\overline{D}$. First we shall prove a preliminary estimation.

THEOREM 3.2. *Let the assumptions of Theorem* 1.1 *and* (2.1*) *hold. Then*

$$\max_{x\in\partial D, 1\le k\le N} \left|\frac{\partial u_k}{\partial x_i}\right| \le c_1, \quad \max_{x\in\overline{D},, 1\le k\le N} \left|\frac{\partial u_k}{\partial x_i}\right| \le \frac{c_2}{\varepsilon}, \tag{3.4}$$

where the constants c_1, c_2 do not depend on h, ε.

Proof. By virtue of (1.5) and (2.4), the first estimate (3.4) will follow from the boundedness of $\frac{\partial F_k}{\partial x_i}$ on ∂D. The functions $F_k(x,h,\varepsilon)$ are the solutions of the problem (1.3), (1.4). By (3.2*) F_k is nonnegative in $\overline{D}$. Then (1.5) implies that

$$\{x \in D : w_k > 1 + \varepsilon\} \subset \{x \in D : u_k > 1 + \varepsilon\}.$$

By (2.4) there exists a positive constant d not depending on h, ε, k, such that

$$\operatorname{dist}(\{x \in D : u_k > 1 + \varepsilon\}, \partial B_2) \ge \operatorname{dist}(\{x \in D : w_k > 1 + \varepsilon\}, \partial B_2) \ge d.$$

We denote

$$w(x) = \lambda \frac{\sqrt{h}}{r} Sh \frac{R_2 - r}{\sqrt{h}}, \ r = |x|.$$

The function $w(x)$ in D satisfies the equation

$$\Delta w - \frac{w}{h} = 0.$$

Hence

$$\Delta(w - F_k) - \frac{w - F_k}{h} = 0 \text{ in } \quad D_d = \{x \in D : R_2 - d < |x| < R_2\}, \quad (w - F_k)|_{\partial D_d} \geq 0,$$

$w - F_k = 0$ on ∂B_2. This easily implies

$$\Big|\frac{\partial F_k}{\partial x_i}\Big|_{\partial B_2} \leq \frac{\lambda}{R_2}.$$

A similar estimate can be proved on ∂B_1 as well.

Furthermore, we differentiate (3.3) with respect to one of the variables x_i and transform it to the following form:

$$\Delta u_k' - [1 - \lambda\chi_\varepsilon'(u_k)]\frac{u_k'}{h} = -\frac{w_{k-1}'}{h} + \frac{\lambda}{h}\chi_\varepsilon'(u_0)u_0'. \tag{3.5}$$

From this relation at the points of a local extremum we obtain the second estimation of (3.4). □

Set

$$\omega_0(\varepsilon) = \{x \in \overline{D} : 1 < u_0(x) = \psi(x) < 1 + \varepsilon\}, \; D_\varepsilon = D \setminus \overline{\omega}_0(\varepsilon).$$

THEOREM 3.3. *Let the assumptions of Theorem* 3.2 *hold and* $\operatorname{dist}(\partial\omega_0, \partial D) \geq h^\sigma$, $\sigma \in (0, 1/2)$, $\varepsilon^4 \geq \sqrt{h}$. *Then there exists a constant* c, *which does not depend on* h, ε, *such that the following estimate holds:*

$$\max_{x \in \overline{D}_\varepsilon, 1 \leq k \leq N} \Big|\frac{\partial u_k}{\partial x_i}\Big| \leq c. \tag{3.6}$$

Proof. First of all, let us prove that $\frac{\partial u_k}{\partial x_i}$ are bounded on the boundary of the domain $\omega_0(\varepsilon)$. Let $x_0 \in \partial\omega_0(\varepsilon)$, $K_R(x_0) \subset D$, $a_0(x_0) = 1 - \lambda\chi_\varepsilon'[u_k(x_0, h, \varepsilon)]$, $a(x) = 1 - \lambda\chi_\varepsilon'[u_k(x, h, \varepsilon)]$,

$$E(|x - y|) = \frac{Sh\left(\sqrt{\frac{a_0(x_0)}{h}}(R - |x - y|)\right)}{4\pi|x - y|Sh\left(\sqrt{\frac{a_0(x_0)}{h}}R\right)}.$$

Let us construct an integral representation of a solution of (3.5).

$$\frac{\partial u_k}{\partial x_i} = \int\limits_{K_R(x_0)} \frac{\partial w_{k-1}}{\partial y_i}\frac{E(|x_0 - y|)}{h}\, dy - \int\limits_{\partial K_R(x_0)} \frac{\partial u_k}{\partial y_i}\frac{\partial E}{\partial n} ds$$

$$+ \int\limits_{K_R(x_0)} [a(x_0) - a(y)]\frac{\partial u_k}{\partial y_i}\frac{E(|x_0 - y|)}{h} dy - \lambda \int\limits_{K_R(x_0)} \chi_\varepsilon'(u_0)\frac{\partial u_0}{\partial y_i}\frac{E(|x_0 - y|)}{h} dy. \tag{3.7}$$

As

$$|a(x_0)-a(y)|\le\frac{c_1}{\varepsilon^3}|x_0-y|,\quad |\chi'_\varepsilon[u_0(y)]|=|\chi'_\varepsilon[u_0(y)]-\chi'_\varepsilon[u_0(x_0)]|$$

$$\le\frac{c_2}{\varepsilon^2}|x_0-y|\max_{x\in\overline{D}}|\nabla u_0|,\quad \frac{\partial E}{\partial n}\Big|_{\partial K_R(x_0)}=-\sqrt{\frac{a_0}{h}}\frac{1}{4\pi RSh\sqrt{\frac{a_0}{h}}R},$$

(3.7) implies

$$\Big|\frac{\partial u_k}{\partial x_i}\Big|\le\max_{y\in\overline{D},1\le k\le N}\Big|\frac{\partial w_{k-1}}{\partial y_i}\Big|\int\limits_{K_R(x_0)}\frac{E(|x_0-y|)}{h}dy+\sqrt{\frac{a_0}{h}}\frac{R}{Sh\sqrt{\frac{a_0}{h}}R}\max_{y\in\overline{D},1\le k\le N}\Big|\frac{\partial u_k}{\partial y_i}\Big|$$

$$+\left(\frac{c_3}{\varepsilon^3}\max_{y\in\overline{D},1\le k\le N}|\nabla u_k|+\max_{y\in\overline{D}}|\nabla u_0|^2\right)\int\limits_{K_R(x_0)}|x_0-y|\frac{E(|x_0-y|)}{h}dy. \tag{3.8}$$

Furthermore,

$$\int\limits_{K_R(x_0)}\frac{E(|x_0-y|)}{h}dy\le 1,\ \int\limits_{K_R(x_0)}|x_0-y|\frac{E(|x_0-y|)}{h}dy\le c_4\sqrt{h}. \tag{3.9}$$

Let us assume that $R\ge h^\sigma,\ \sigma\in(0,1/2)$. Hence, by the known inequality

$$x^m\exp(-x)\le m^m\exp(-m)\quad\forall x\ge 0,\ m>0, \tag{3.10}$$

we get

$$\sqrt{\frac{a_0}{h}}R\frac{1}{Sh\sqrt{\frac{a_0}{h}}R}\le ch^{\sigma_1},\quad \sigma_1>0. \tag{3.11}$$

From (3.8), taking into account (3.5), (3.9)–(3.11), we obtain

$$\Big|\frac{\partial u_k}{\partial x_i}\Big|\le\max_{x\in\overline{D},1\le k\le N}|\nabla w_{k-1}|+\frac{c}{\varepsilon}h^{\sigma_1}+\frac{c}{\varepsilon^4}\sqrt{h}+\frac{c}{\varepsilon^2}\sqrt{h}\max_{x\in\overline{D}}|\nabla u_0|^2.$$

From here, assuming that $\varepsilon^4\ge\sqrt{h}$, we obtain estimation (3.6) on the boundary of the domain $\omega_0(\varepsilon)$. On the boundary ∂D estimation (3.6) is proved in Theorem 3.2. To complete the proof it suffices to notice that in the domain $D_\varepsilon=D\setminus\overline{\omega}_0(\varepsilon)$ (3.5) has the following form:

$$\Delta u'_k-[1-\lambda\chi'_\varepsilon(u_k)]\frac{u'_k}{h}=-\frac{w'_{k-1}}{h}.$$

Therefore, if $u'_k(x,h,\varepsilon)$ attains its own extremum in D_ε, the estimation (3.6) is obvious. □

Corollary 3.1. *Let the assumptions of Theorem* 3.3 *hold and*

$$x_0\in\{x\in D:\psi(x)\ge 1+\varepsilon\},\ \ y_0\in\{x\in D:\psi(x)\le 1\}.$$

Then

$$\left|\frac{\partial u_k(x_0,h,\varepsilon)}{\partial x_i}-\frac{\partial u_k(y_0,h,\varepsilon)}{\partial x_i}\right|\le c_1|x_0-y_0|^\alpha+c_2h^{\sigma_1},$$

$$|u_k(x_0,h,\varepsilon)-u_k(y_0,h,\varepsilon)|\le |x_0-y_0|+c_2h^{\sigma_1}, \tag{3.11*}$$

where $c_i>0,\sigma_1>0$ do not depend on h,ε,k,x_0,y_0.

The proof of this statement can be carried out by the same techniques as that of Theorem 3.3.

Set $\omega_k(h,\varepsilon)=\{x\in D: 1<u_k<1+\varepsilon\}$.

THEOREM 3.4. *Let the assumptions of Theorem* 3.1 *hold. If $x\in\omega_k^+=\{x\in D: u_k(x,h,\varepsilon)\ge 1+\varepsilon\}$ and $\operatorname{dist}(x,\partial\omega_k^+)\ge h^\sigma$, $\sigma\in(0,1/2)$, then*

$$\max_x|F_k(x,h,\varepsilon)|\le c_1h^{\sigma_1},\quad \sigma_1>0. \tag{3.12}$$

If $x\in D_k(h,\varepsilon)=D\setminus[\omega_k(h,\varepsilon)\cup\omega_0(h,\varepsilon)]$ and $\operatorname{dist}(x,\partial D_k)\ge h^\sigma$, then

$$\max_x\Big|\frac{\partial^\beta F_k}{\partial x_1^{\beta_1}\partial x_2^{\beta_2}\partial x_3^{\beta_3}}\Big|\le\lambda c_2(\beta h^{\sigma_1})^\beta,\quad \sigma_1>0,\ \beta=\beta_1+\beta_2+\beta_3. \tag{3.13}$$

If $x\in D\setminus[\overline{\omega}_k(h,\varepsilon)\cup\overline{\omega}_{k-1}(h,\varepsilon)]$ and $\operatorname{dist}\Big(x,\partial\{D\setminus[\overline{\omega}_k(h,\varepsilon)\cup\overline{\omega}_{k-1}(h,\varepsilon)]\}\Big)\ge h^\sigma$, then

$$\max_x\Big|\frac{F_k-F_{k-1}}{h}\Big|\le c_3h^{\sigma_1},\quad \sigma_1>0. \tag{3.14}$$

The constants c_i,σ_1,β do not depend on k, h, ε.

Proof. From (1.3), (1.4) it follows that

$$\Big|F_k(x,h,\varepsilon)\Big|\le\lambda\quad\forall x\in\overline{D}.$$

Let $x_0\in\omega_k^+$ and $\operatorname{dist}(x_0,\partial\omega_k^+)\ge h^\sigma$, $R\ge h^\sigma$, $\sigma\in(0,1/2)$. Then

$$|F_k(x_0,h,\varepsilon)|\ \le\ -\int\limits_{\partial K_R(x_0)}|F_k|\frac{\partial E(|x_0-y|)}{\partial n}ds\le\lambda\sqrt{\frac{1}{h}}R\frac{1}{Sh\sqrt{\frac{1}{h}}R},$$

where

$$E(|x_0-y|)=\frac{Sh\sqrt{\frac{1}{h}}(R-|x_0-y|)}{4\pi|x_0-y|Sh\sqrt{\frac{1}{h}}R}.$$

After that (3.14) follows from (3.11).

Let us differentiate (1.3) with respect to x_i. It gives

$$\Delta\frac{\partial F_k}{\partial x_i}-\frac{1}{h}\frac{\partial F_k}{\partial x_i}=-\frac{\lambda}{h}\chi_\varepsilon'(u_k)\frac{\partial u_k}{\partial x_i}+\frac{\lambda}{h}\chi_\varepsilon'(u_0)\frac{\partial u_0}{\partial x_i}.$$

On $D_k(h,\varepsilon)$ this equation takes the form

$$\Delta\frac{\partial F_k}{\partial x_i}-\frac{1}{h}\frac{\partial F_k}{\partial x_i}=0.$$

Let $\operatorname{dist}(x_0,\partial D_k(h,\varepsilon))\ge h^\sigma$, $K_R(x_0)\subset D_k(h,\varepsilon)$. Then we have

$$\frac{\partial F_k(x_0,h,\varepsilon)}{\partial x_i}=-\int\limits_{\partial K_R(x_0)}\frac{\partial F_k}{\partial y_i}\frac{\partial E(|x_0-y|)}{\partial n}ds.$$

Taking into account that $\frac{\partial}{\partial n}E(|x_0-y|)$ on $\partial K_R(x_0)$ does not depend on the variable of integration, we divide by it both sides of the previous equality and integrate over R. We get

$$\frac{\partial F_k(x_0,h,\varepsilon)}{\partial x_i}\int\limits_0^R 4\pi\sqrt{h}\rho Sh\sqrt{\frac{1}{h}}\rho d\rho = -\int\limits_{K_R(x_0)}\frac{\partial F_k}{\partial y_i}dy.$$

As

$$\int\limits_0^R \sqrt{h}\rho Sh\frac{\rho}{\sqrt{h}}d\rho = RhCh\sqrt{\frac{1}{h}}R - h^{\frac{3}{2}}Sh\sqrt{\frac{1}{h}}R$$

and $R \geq h^\sigma$, using the inequality (3.11), we obtain

$$\Big|\frac{\partial F_k}{\partial x_i}\Big| \leq c_m h^{m(\frac{1}{2}-\sigma)-\frac{3}{2}}, \quad m>0.$$

We construct a sequence of domains $L_1, L_2, \ldots, L_{l+1}$ such that

$$L_{l+1} = D \setminus [\overline{\omega}_k(h,\varepsilon)\cup\overline{\omega}_0(\varepsilon)], L_j \subset L_{j+1},$$

and the distance between L_j and ∂L_{j+1} equals $\frac{h^\sigma}{l}$, $j=1,2,\ldots,l$. For any point x_0 in L_j the ball with the radius $\frac{h^\sigma}{l}$ and center x_0 is contained in L_{j+1} : $K_{\frac{h^\sigma}{l}}(x_0)\subset L_{j+1}$. Hence, applying the last inequality to the equation

$$\Delta D^{l+1-j}F_k - \frac{1}{h}D^{l+1-j}F_k = 0 \ \text{ in } L_j, \ \ D^j = \frac{\partial^j}{\partial x_1^{n_1}\partial x_2^{n_2}\partial x_3^{n_3}}, j = n_1+n_2+n_3,$$

we get

$$\max_{\overline{L}_j}|D^{l+1-j}F_k(x,h,\varepsilon)| \leq c_m h^{\frac{m}{l}(\sigma-\frac{1}{2})-\frac{3}{2}}\max_{\overline{L}_{j+1}}|D^{l-j}F_k(x,h,\varepsilon)|.$$

By considering these inequalites one by one for $j=1,2,\ldots,l$ and estimating the right-hand side of the jth inequality using the $(j+1)$th inequality, we finally get

$$\max_{x\in\overline{D}_k(h,\varepsilon}|D^lF_k(x,h,\varepsilon)| \leq [c_m h^{\frac{m}{l}(\sigma-\frac{1}{2})-\frac{3}{2}}]^l\max_{x\in\overline{D}}|F_k(x,h,\varepsilon)|.$$

Since m is an arbitrary positive number, we can choose it so that $\sigma_1 = m(\sigma-\frac{1}{2})-\frac{3}{2}l > 0$. The proof of (3.14) is quite similar. □

Denote

$$\begin{aligned}\omega_k(h,\varepsilon,\sigma) &= \{x\in D: 1-h^\sigma < u_k(x,h,\varepsilon) < 1+\varepsilon+h^\sigma,\}, \ \omega_k^1(h,\varepsilon,\sigma)\\ &= \{x\in D: 1-h^\sigma < u_{k-1}(x,h,\varepsilon) < 1+\varepsilon+h^\sigma\}\cup\omega_k(h,\varepsilon,\sigma), \ \omega_0(h,\varepsilon,\sigma)\\ &= \{x\in D: 1-h^\sigma < u_0 < 1+\varepsilon+h^\sigma\}.\end{aligned}$$

THEOREM 3.5. *Let the assumptions of Theorem* 3.2 *hold and*

$$\Big\|\frac{\varphi_k-\varphi_{k-1}}{h}\Big\|_{C^\alpha(\overline{D})} + \|\varphi_k\|_{C^{2+\alpha}(\overline{D})} \leq c.$$

Then $\forall h > 0, \varepsilon^3 \geq \sqrt{h}$ *the following estimations hold:*

$$\left\| \frac{u_k - u_{k-1}}{h} \right\|_{C^\alpha(\overline{\Omega}_k^1)} + \|u_k\|_{C^{2+\alpha}(\overline{\Omega}_k)} \leq \|w_k\|_{C^{2+\alpha}(\overline{D})} + c_1 h^{\sigma_1}, \sigma_1 > 0, \tag{3.15}$$

$$\Omega_k = \{x \in D \setminus [\overline{\omega}_k(h,\varepsilon,\sigma) \cup \overline{\omega}_0(h,\varepsilon)] : \ \operatorname{dist}(x, \partial D) \geq h^\sigma\},$$

$$\Omega_k^1 = \{x \in D \setminus [\overline{\omega}_k^1(h,\varepsilon,\sigma) \cup \overline{\omega}_0(h,\varepsilon)] : \ \operatorname{dist}(x, \partial D) \geq h^\sigma\},$$

$$\min_{x \in \overline{\Omega}_k^1} \frac{u_{k-1} - u_k}{h} \geq \min_{x \in \overline{D}} \frac{w_{k-1} - w_k}{h} - c_2 h^{\sigma_2}, \quad \sigma_2 > 0, \tag{3.16}$$

the constants c_i, σ_i *do not depend on* h, ε, *and* k.

Proof. Note that (3.17) and (3.18) follow from (3.13) and (3.14). After that the statements of the theorem become obvious. □

THEOREM 3.6. *Let* $\varphi_1(x,t) \equiv 0$, $\varphi_2(x,t) \equiv q = \text{const}$, $q > 1$,

$$\psi \in C^{2+\alpha}(\overline{D}) \ \ \frac{\partial \psi}{\partial \rho} > 0 \ in \ \overline{D}, \ \rho = |x|, \ \ \Delta\psi \leq 0 \quad in \quad \overline{D}, \tag{3.17}$$

then

$$\min_{x \in \overline{D}, 1 \leq k \leq N} \frac{\partial w_k}{\partial \rho} \geq c_2, \quad \min_{x \in \overline{\Omega}_k, 1 \leq k \leq N} \frac{\partial u_k}{\partial \rho} \geq \min_{x \in \overline{D}, 1 \leq k \leq N} \frac{\partial w_k}{\partial \rho} - c_3 h^{\sigma_1}, \quad \sigma_1 > 0, \tag{3.18}$$

where the constants $c_i, i = 1,2,3$, σ_1 *do not depend on* h, ε, *and* k.

Proof. As Theorem 2.1 implies,

$$\Delta w_k = \frac{w_k - w_{k-1}}{h} \leq 0.$$

Let $v(x)$ be a function which satisfies the following conditions:

$$\Delta v = 0 \ \text{ in } \ D, \ \ v = 0 \ \text{ on } \ \partial B_1, \ \ v = q \ \text{ on } \ \partial B_2.$$

Obviously,

$$v(x) \leq w_k(x,h,\varepsilon) \leq w_0 = \psi(x) \ \text{ in } \ D.$$

This implies

$$\left.\frac{\partial w_k}{\partial \rho}\right|_{\partial B_2} \geq \left.\frac{\partial \psi}{\partial \rho}\right|_{\partial B_2}, \ \ \left.\frac{\partial w_k}{\partial \rho}\right|_{\partial B_1} \geq \left.\frac{\partial v}{\partial \rho}\right|_{\partial B_1}.$$

Assume now that

$$\min_{x \in \overline{D}, 1 \leq k \leq N} \left(\rho \frac{\partial w_k}{\partial \rho}\right) = \left(\rho \frac{\partial w_n}{\partial \rho}\right)\Big|_{x = x_0}, \quad x_0 \in D.$$

We differentiate (1.6) with respect to ρ. It will give

$$\Delta\left(\rho \frac{\partial w_n}{\partial \rho}\right) - \frac{1}{h}\left(\rho \frac{\partial w_n}{\partial \rho} - \rho \frac{\partial w_{n-1}}{\partial \rho}\right) = 2\frac{w_n - w_{n-1}}{h}.$$

As at a point of local minimum of $\left(\rho\frac{\partial w_n}{\partial\rho}\right)$, $\Delta(\rho\frac{\partial w_n}{\partial\rho})\le 0$, we have

$$\rho\frac{\partial w_n}{\partial\rho}-\rho\frac{\partial w_{n-1}}{\partial\rho}\ge 2(w_{n-1}-w_n)\ge 0.$$

Hence, there exists a constant $c_1>0$ such that it does not depend on h, ε, and k, and

$$\min_{x\in\overline{D},1\le k\le N}\frac{\partial w_k}{\partial\rho}\ge c_1>0.$$

By (1.5) we get

$$\frac{\partial u_k}{\partial\rho}=\frac{\partial w_k}{\partial\rho}+\frac{\partial F_k}{\partial\rho}.$$

The relation (3.13) completes the proof of (3.20). □

4. Passage to the limit. Let the function $\eta(x,t)\in C^{2,1}(\overline{D})$ be equal to zero on $(\partial D\times(0,T))\cup(D\times(t=T))$, $\eta_k(x)=\eta(x,kh)$. Let us represent (1.1) as

$$\Delta u_k-\frac{u_k-u_{k-1}}{h}=-\frac{\lambda}{h}[\chi_\varepsilon(u_k)-\chi_\varepsilon(u_{k-1})]+\Delta F_{k-1}.$$

We multiply (1.1) by $h\eta_k(x)$, integrate it over D, and take the sum over k from 1 to N. After simple transformations we obtain

$$h\sum_{k=1}^{N}\int_D\Big\{\nabla u_k\nabla\eta_k+\frac{1}{h}(u_k-u_{k-1})\eta_k+\lambda\chi_\varepsilon(u_k)\frac{\eta_{k+1}-\eta_k}{h}\Big\}dy$$

$$+\lambda\int_D\chi_\varepsilon(u_0)\eta_1dy+h\sum_{k=1}^{N-1}\int_D\frac{F_k-F_{k-1}}{h}\psi_{k+1}dy=0, \tag{4.1}$$

where $\psi_k=h\sum\limits_{l=k}^{N}\Delta\eta_l$. Let us denote by $\{\overline{u}(x,t,h,\varepsilon)\}$ the piecewise linear interpolations of the functions $\{u_k(x,h,\varepsilon)\}$ with respect to the variable t. If $t\in[(k-1)h,kh]$, then

$$\overline{u}(x,t,h,\varepsilon)=u_{k-1}(x,h,\varepsilon)+\frac{u_k(x,h,\varepsilon)-u_{k-1}(x,h,\varepsilon)}{h}(t-(k-1)h).$$

Let

$$\begin{aligned}G_T(h,\varepsilon)&=\{(x,t)\in D_T:\overline{u}(x,t,h,\varepsilon)>1+\varepsilon+h^\sigma\},\ \ \Omega_T(h,\varepsilon)\\&=\{(x,t)\in D_T:\overline{u}(x,t,h,\varepsilon)<1-h^\sigma\},\ \ \gamma_T^-(h,\varepsilon)=\partial\Omega_T\cap D_T,\\\gamma_T^+(h,\varepsilon)&=\partial G_T\cap D_T,\\\gamma_0^+(\varepsilon)&=\{x\in D:\psi(x)=1+\varepsilon\}.\end{aligned}$$

THEOREM 4.1. *Let the following conditions be satisfied:*

$$\psi(x)\in C^{2+\alpha}(\overline{D}),\quad \Delta\psi\le 0\quad in\quad\overline{D},\quad \frac{\partial\psi}{\partial\rho}>0\quad in\quad\overline{D},$$

$$\varphi_1(x,t)=0 \quad on \quad \partial B_1, \quad \varphi_2(x,t)=q=\text{const}>1 \quad on \quad \partial B_2,$$

and assume that the corresponding compatibility conditions at $t=0$, $x\in\partial\Omega_T\cup\partial G_T$ *hold. Then* $\forall\, T>0$, *there exists a unique solution of the problem* (0.1)–(0.4) *and*

$$u(x,t)\in C(\overline{D}_T)\cap\Big(H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{\Omega}_T)\times H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{G}_T)\Big);$$

the free boundary is given by the graph $\rho=\omega(\theta_1,\theta_2,t)$ *of a function* $\omega(\theta_1,\theta_2,t)\in H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{\Pi}_T)$, *where* (ρ,θ_1,θ_2) *are spherical coordinates,* $\Pi=\{(\theta_1,\theta_2):0<\theta_1<2\pi, 0<\theta_2<\pi\}$, $\Pi_T=\Pi\times(0,T)$.

Proof. From the estimations (3.2*) and (3.6) it follows that $\dfrac{\partial\overline{u}}{\partial t}$ is uniformly bounded everywhere in $\overline{D}_T$ and $|\nabla\overline{u}|$ is uniformly bounded everywhere in $\overline{D}_T$ excluding the set, whose measure vanishes as $h,\varepsilon\to 0$.

Denote

$$u(x,t)=\lim_{\varepsilon,h\to 0}\overline{u}(x,t,h,\varepsilon), \quad \varepsilon^4\geq\sqrt{h}.$$

From (3.20) it follows that

$$\frac{\partial\overline{u}}{\partial\rho}\geq c>0 \quad \text{on} \quad \Omega_T(h,\varepsilon)\cup G_T(h,\varepsilon). \tag{4.2}$$

Therefore, the level surfaces $\gamma_T^{\pm}(h,\varepsilon)$ can be given by the explicit equations

$$\rho=\omega^{+}(\theta,t,h,\varepsilon), \quad \rho=\omega^{-}(\theta,t,h,\varepsilon),$$

respectively, and

$$\omega_t^{\pm}=-\frac{\overline{u}_t^{\pm}}{\overline{u}_\rho^{\pm}}, \quad \omega_{\theta_i}^{\pm}=-\frac{\overline{u}_{\theta_i}^{\pm}}{\overline{u}_\rho^{\pm}}, \quad i=1,2,$$

are uniformly bounded. The functions $\frac{1}{h}[F_k(x,h,\varepsilon)-F_{k-1}(x,h,\varepsilon)]$ are nonnegative and uniformly bounded in $\overline{D}$, and, besides, the estimate (3.14) holds.

The assumptions of Theorem 4.1 on the initial conditions imply

$$\lim_{\varepsilon\to 0}\text{mes}\{1\leq\psi\leq 1+\varepsilon\}=0, \quad \lim_{\varepsilon\to 0}\gamma_0^{+}(\varepsilon)=\gamma_0,$$

and Corollary 3.1 implies that for $\overline{u}(x,t,h,\varepsilon)$, (3.11*) holds.

The facts stated above allow us to pass to the limit in (4.1) as $\varepsilon\to 0, h\to 0, \varepsilon^4\geq\sqrt{h}$. It gives

$$\int\limits_{D_T}\Big[\nabla u\nabla\eta+u_t\eta\Big]dxdt+\lambda\int\limits_{D_T\cap\{u<1\}}\eta_t dxdt+\lambda\int\limits_{\Omega_0}\eta(x,0)dx$$
$$+\lambda\int\limits_{D_T^1}\lim_{\varepsilon,h\to 0}\chi_\varepsilon(\overline{u})\eta_t dxdt+\int\limits_{D_T^1}\lim_{\varepsilon,h\to 0}F_t\int\limits_t^T\Delta\eta d\tau\, dxdt=0,$$

where $D_T^1=D_T\cap\{u=1\}$. This integral identity implies that

$$\Delta u-u_t=0 \ \text{ in } \ \Omega_T\cup G_T,$$

where $\Omega_T = \lim\limits_{\varepsilon,h\to 0} \Omega_T(h,\varepsilon)$, $G_T = \lim\limits_{\varepsilon,h\to 0} G_T(h,\varepsilon)$,

$$\lambda \int\limits_{D_T^1} \lim_{\varepsilon,h\to 0} \chi_\varepsilon(\overline{u})\eta_t dxdt + \int\limits_{D_T^1} \lim_{\varepsilon,h\to 0} F_t \int\limits_t^T \Delta\eta(x,\tau)d\tau dxdt$$
$$+ \int\limits_{\gamma_T^-} [\frac{\partial u}{\partial n^-} + \lambda\cos(n^-,t)]\eta ds + \int\limits_{\gamma_T^+} \frac{\partial u}{\partial n^+}\eta ds = 0,$$

where $n^\pm$ are the outward pointing normal vectors to $\partial\Omega_T, \partial G_T$, respectively. Let $\eta(x,t) = 0$ on γ_T^-. Then

$$\lambda \int\limits_{D_T^1} \lim_{\varepsilon,h\to 0} \chi_\varepsilon(\overline{u})\eta_t dxdt + \int\limits_{D_T^1} \lim_{\varepsilon,h\to 0} F_t \int\limits_t^T \Delta\eta(x,\tau)d\tau dxdt + \int\limits_{\gamma_T^+} \frac{\partial u}{\partial n^+}\eta ds = 0. \tag{4.3}$$

Let us take $\eta(x,t)$ as a solution of the following problem:

$$\begin{gathered} \Delta\eta - \eta_t = -f_t(x,t) \text{ in } D_T \setminus \overline{\Omega}_T, \\ \eta(x,t)\Big|_{\gamma_T^-} = 0, \ \ \eta(x,t)\Big|_{\partial D_2} = 0, \ \ \eta(x,0) = \Phi(x) \geq 0 \text{ in } D \setminus \Omega_0, \end{gathered} \tag{4.4}$$

where $f(x,t)$, $\Phi(x)$ are given smooth functions. The change of variables

$$\rho' = R_1 + \frac{\rho - \omega^-(\theta,t)}{R_2 - \omega^-(\theta,t)}, \ \theta' = \theta, \ t' = t$$

maps the domain $D_T \setminus \overline{\Omega}_T$ onto a cylinder, and (4.5) in new coordinates will have varying coefficients. The solvability of this problem follows from [11, chap. 4, §9]. Let us choose $f(x,t), \Phi(x)$ so that

$$f_t(x,t) \geq 0, f_{tt}(x,t) \leq 0, f(x,T) = 0, \Delta\Phi + f_t(x,0) \leq 0.$$

Then the maximum principle implies that

$$\eta(x,t) \geq 0, \ \ \eta_t(x,t) \leq 0.$$

Note also that $\frac{\partial u}{\partial n^+} \leq 0$ on γ_T^+. All these facts and the identity (4.4) imply

$$\int\limits_{D_T^1} \lim_{\varepsilon,h\to 0} F_t f(x,t)dxdt \geq 0.$$

Hence, $\text{mes}\,\{(x,t) \in D_T^1 : \lim\limits_{\varepsilon,h\to 0} F_t > 0\} = 0$. After that one can easily prove that $\text{mes}\,\{(x,t) \in D_T^1 : \lim\limits_{\varepsilon,h\to 0} \chi_\varepsilon(\overline{u}) > 0\} = 0$.

As a result we obtain that

$$\int\limits_{D_T} \Big\{\nabla u\nabla\eta + u_t\eta + \lambda\chi(u)\eta_t\Big\}dxdt + \lambda \int\limits_D \chi(\psi)\eta(x,0)dx = 0, \tag{4.5}$$

where $\chi(u)$ is the characteristic function of the domain Ω_T, $\chi(\psi)$ is the characteristic function of the domain Ω_0. Identity (4.6) implies $\gamma_T^+ = \gamma_T^- = \gamma_T = \partial\Omega_T \cap D = \partial G_T \cap D$, and the equation of the free boundary $\gamma_T = \partial\Omega_T \cap D = \partial G_T \cap D$ has the form $\rho = \omega(\theta, t)$, where ω_θ, ω_t are uniformly bounded. As follows from the estimation (3.17), $\|u_t(x,t)\|_{C^\alpha(\overline{\Omega}_T \cup \overline{G}_T)}$, $\|u_{xx}(x,t)\|_{C^\alpha(\overline{\Omega}_T \cup \overline{G}_T)}$ are bounded by a constant which does not depend on t. The domains obtained by cutting the domains Ω_T, G_T by planes $t = \text{const.}$ have the cone property. Then, by [21, chap. 2, §3] we get

$$u_x(x,t) \in H^{1,\alpha/2}(\overline{\Omega}_T) \times H^{1,\alpha/2}(\overline{G}_T).$$

Then (4.6) implies that the condition (0.3) holds everywhere on the free boundary. From this we obtain that

$$u(x,t) \in \left\{ H^{2+\alpha.1+\alpha/2}(\overline{\Omega}_T) \times H^{2+\alpha,1+\alpha/2}(\overline{G}_T) \right\} \cap C(\overline{D}_T), \frac{\partial u^\pm}{\partial \rho}\Big|_{\gamma_T} \geq c > 0,$$

the free boundary is given by the graph $\rho = \omega(\theta_1, \theta_2, t)$ of a function $\omega(\theta_1, \theta_2, t)$ $\in H^{2+\alpha,1+\frac{\alpha}{2}}(\overline{\Pi}_T)$, where $(\rho, \theta_1, \theta_2)$ are spherical coordinates, $\Pi = \{(\theta_1, \theta_2) : 0 < \theta_1 < 2\pi, 0 < \theta_2 < \pi\}$, $\Pi_T = \Pi \times (0, T)$.

The uniqueness of solution of the Stefan problem is proved, for example, in [21, chap. 5, §9]. □

Remark. If in the hypothesis of the theorem we omit the requirement $\psi_\rho > 0$ in $\overline{D}$, then the existence of global classical solution can be proved if we additionally assume that

$$\Delta\psi < 0 \quad \text{in} \quad \overline{D}, \quad \frac{\partial \varphi_i}{\partial t} < 0 \quad \text{in} \quad \overline{D}_T.$$

Then, the free boundary is the graph of a function of $C^{1+\frac{\alpha}{2}}$ class.

REFERENCES

[1] O. A. Oleinik, *On a method of solution of a general Stefan problem*, Dokl. Akad. Nauk USSR, 135 (1960), pp. 1054–1057.

[2] S. Kamenomostkaya, *On the Stefan problem*, Math. Sb., 5 (1961), pp. 489–514.

[3] C. Baiocchi, *Sur un problème à frontière libre traduisant le filtrage de liquides atraverse des milieux poreux*, C.R. Acad. Sci. Paris Sér. A, 273 (1971), pp. 1215–1217.

[4] G. Duvaut, *Résolution d'un problème de Stefan*, C.R. Acad. Sci. Paris Sér. A, 276 (1973), pp. A1461–A1463.

[5] A. Friedman and D. Kinderlehrer, *A one phase Stefan problem*, Indiana Univ. Math. J., 24 (1975), pp. 1005–1035.

[6] D. Kinderlehrer and L. Nirenberg, *The smoothness of the free boundary in the one phase Stefan problem*, Comm. Pure Appl. Math., 31 (1978), pp. 257–282.

[7] L. A. Caffarelli, *The smoothness of the free surface in a filtration problem*, Arch. Rational Mech. Anal., 63 (1976), pp. 77–86.

[8] L. A. Caffarelli, *The regularity of elliptic and parabolic free boundaries*, Bull. Amer. Math. Soc., 82 (1976), pp. 616–618.

[9] L. A. Caffarelli, *The regularity of free boundaries in higher dimensions*, Acta Math., 139 (1977), pp. 156–184.

[10] M. Fremon, *Variational formulation of the Stefan problem. Frost propagation in porous media*, in International Conference on Computational Methods in Nonlinear Mechanics, Austin, Texas, 1974, pp. 341–350.

[11] R. H. Nochetto, *A class of nondegenerate two-phase Stefan problems in several space variables*, Comm. Partial Differential Equations, 12 (1987), pp. 21–45.

[12] J. Athanasopoulos, L. Caffarelli, and S. Salsa, *Regularity of the free boundary in parabolic phase transition problems*, Acta Math., 176 (1996), pp. 245–282.

[13] A. M. Meirmanov, *The Stefan Problem*, Novosibirsk, Nauka, 1986.
[14] B. V. Basaliy, *The Stefan problem*, Dokl. Akad. Nauk Ukrain., Ser. A, 1986, pp. 3–7.
[15] E. V. Radkevich and A. K. Melikulov, *Boundary Value Problems with Free Boundary*, FAN, Tashkent, 1988.
[16] M. A. Borodin, *Existense of the classic solution of a two-phase multidimensional Stefan problem on any finite time interval*, Intern. Ser. Numer. Math., 106 (1992), pp. 97–103.
[17] M. Meirmanov, *An example of the nonexistence of classic solution for a Stefan problem*, Dokl. Akad. Nauk USSR, 258 (1981), pp. 547–549.
[18] M. Primicherio, *Mashy regions in phase-change problems*, in Applied Nonlinear Functional Analysis, Lang, Franfurt, Main, Germany, 1982, pp. 251–269.
[19] M. A. Borodin, *A two-phase contact Stefan problem*, Ukranian Math. J., 47 (1995), pp. 158–167.
[20] O. A. Ladyshenskaja and N. N. Uraltceva, *Linear and Quasilinear Elliptic Equations*, Nauka, Moscow, 1973.
[21] O. A. Ladyshenskaja, V. A.Solonnikov, and N. N. Uraltceva, *Linear and Quasilinear Parabolic Equations*, Nauka, Moscow, 1967.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1282–1308

A MATHEMATICAL STUDY OF THE RELAXED OPTICAL FLOW PROBLEM IN THE SPACE $BV(\Omega)$*

G. AUBERT† AND P. KORNPROBST†

Abstract. This paper describes a variational approach for estimating a discontinuous optical flow from a sequence of images. Defined as the apparent motion of the image brightness pattern, the optical flow is very important in the computer vision community, where its accurate estimation is strongly needed. After a short overview of existing methods, we present a new variational method that we study in the space of bounded variations. We first present an integral representation of the optical flow problem which appears to be not lower semicontinuous. The relaxed functional is then calculated. We conclude by challenging questions about the possible numerical analysis of the abstract results.

Key words. measure theory, space of bounded variations, convex functions of measures, Γ-convergence, elliptic equations, relaxation of ill-posed problems, optical flow, computer vision

AMS subject classifications. 35J, 49J, 65N

PII. S003614109834123X

1. Introduction.

This paper deals with the estimation of the movement in a sequence of images. This velocity field will be called the optical flow. In the computer vision community, it is well known that the optical flow is a rich source of information about the geometrical structure of the world. Many numerical algorithms on the optical flow estimation and its applications have been performed. They have clearly shown how the optical flow can be used to recover information about slant and tilt of surface elements, ego-motion, shape information, time to collision, etc. [31, 32, 30, 34, 33, 29, 28, 38, 50, 24, 37, 49, 27, 9, 40, 41].

Almost all of these approaches use the classical brightness constancy assumption that relates the gradient of brightness to the components of the local flow to estimate. Because this problem is ill-posed, additional constraints are usually required. The most common approach is to add a quadratic smoothness constraint, as done originally by Horn and Schunk [29]. However, in order to estimate the optical flow more accurately, other constraints involving high order spatial derivatives have also been used [40]. Nevertheless, several of the proposed methods lacked robustness to the presence of occlusion and yielded smooth optical flow. The variational approach proposed in this paper is motivated by the need to recover the optical flow while preventing the method from trying to smooth the solution across the flow discontinuities. To cope with discontinuities, we propose in this article a complete mathematical study of the relaxed optical flow problem in the space $BV(\Omega)$. We first present an integral representation of the optical flow problem which does not appear to be lower semicontinuous.

This article is organized as follows.

In section 1, we define the problem and propose a variational approach to solve it. The general idea is based on a conservation law of the intensity along the trajectories. We will deal with an ill-posed problem that we will solve by regularizing the unknowns.

*Received by the editors June 29, 1998; accepted for publication (in revised form) February 26, 1999; published electronically October 8, 1999.
http://www.siam.org/journals/sima/30-6/34123.html
†Laboratoire J.A. Dieudonné, UNSA, UMR 6621 du CNRS, 06108 Nice Cedex 2, France (gaubert@math.unice.fr, pkornp@sophia.inria.fr).

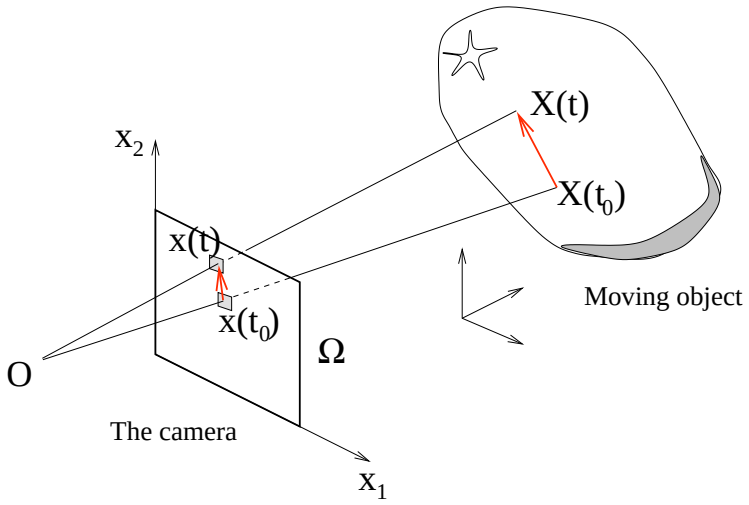


FIG. 2.1. *Pinhole model of camera. Ω is the domain of the image; O is the optical center.*

Section 3 presents some general recalls about the space of bounded variations (noted $BV(\Omega)$). Classically used for problems coming from computer vision, this space permits discontinuities along curves (in dimension 2).

In section 4, we concentrate on the meaning of the energy we defined. This will permit us to consider some integral representation results of the duality pairing of an integrable function with a measure. Similar results have been proved by Anzellotti [3], and we will extend them under weaker assumptions. This will enable us to obtain a fully developed expression for energy that we have to minimize. Unfortunately, the proposed energy is not lower semicontinuous for the weak topology of $BV(\Omega)$.

Section 5 is devoted to the computation of the relaxed functional. This part is mainly technical and is based on ideas developed by Bouchitté et al. [14, 13].

Finally, we prove in section 6 that there exists a solution in $\mathbf{BV}(\Omega)$ for the relaxed formulation.

2. The optical flow problem: Definition and modelization.

2.1. Definition. As shown in Figure 2.1, we can modelize a camera as a simple projective model. Consequently, the first idea is to say that the two-dimensional velocity field in the image corresponds to the projection of the three-dimensional velocity field of the objects. However, variations of intensity due to shadows do not correspond to any real motion. The importance of the light source can be seen toward other phenomena. For instance, if the object is sparkling, the reflected luminosity changes rapidly with the position. This is the case for bodywork, glasses, etc. Finally, notice the unavoidable problem of noise in images. These intensity variations may be interpreted as false motions which have no physical meaning.

Thanks to these remarks, we will define the optical flow as the two-dimensional velocity field describing the changes in intensity between images. In many cases, it can be interpreted as an approximation of the projection of the three-dimensional velocity field which animates physical objects. We will see in the next section how we can traduce it mathematically.

2.2. A short overview. In this last decade, numerous methods have been proposed to compute optical flow. Several ideas have been used: working with regions, curves, lines, or points. There is also a wide range of methodologies: wavelets, Markov random fields, Fourier analysis, and naturally partial differential equations [29, 28, 38, 50, 24, 37, 49, 27, 18, 40, 41]. We refer the interested reader to two (mainly computational) general surveys:

• Barron, Fleet, and Beauchemin [9] explain the main techniques and perform numerical quantitative experiments to compare them. (The database used for tests is also available.)

• Orkisz and Clarysse [39] propose an updated version of the preceding one.

In this article we will concentrate upon the class of *differential methods* (as named by Barron, Fleet, and Beauchemin) which have been proved to be among the best [9]. Their common point is the consistency intensity hypothesis of a point during its movement. More precisely, we will assume that

(2.1) *"The intensity of a point keeps constant along its trajectory."*

This hypothesis is called the *optical flow constraint* (OFC) hypothesis. We can consider it as reasonable, almost along short times, for which changes of the brightness are weak.

Let $x(t) = (x_1(t), x_2(t)) \in \Omega \subset R^2$ be the projection of the point $X(t) \in R^3$ at time t (see Figure 2.1,). For $x \in \Omega$, we denote by $u(t,x)$ the reflected intensity (the brightness) of the point x at time t. Let t_0 be fixed. Using these notations, a natural way to express (2.1) is

$$u(t, x(t)) = u(t_0, x(t_0)). \tag{2.2}$$

By differentiating (2.2) with respect to t, we obtain, for $t = t_0$,

$$\sigma(x) \cdot Du(t_0, x) + u_t(t_0, x) = 0, \qquad x \in \Omega,$$

where $\sigma = (\sigma_1, \sigma_2)^T = \left(\frac{dx_1}{dt}, \frac{dx_2}{dt}\right)^T$ is the unknown velocity field, $D\cdot$ is the spatial gradient operator, and u_t denotes the temporal derivative of $u(t,x)$. (Derivatives are written in the distributional sense.) This equation is the OFC. Naturally, this scalar equation is insufficient to compute both components of the flow field. This problem is usually called the *aperture problem*. Additional constraints are therefore required to reduce the space of admissible functions. Several possibilities are then possible: use additional constraints, consider special movements (rigid or fluids), regularize the velocity field, etc. We refer to [6], where we notably propose an overview of these different methods.

Our starting point will be the method proposed by Horn and Schunk in 1981 [29]. The idea is to minimize the following energy:

$$E_{HS}(\sigma) = \underbrace{\int_\Omega ((\sigma \cdot Du) + u_t)^2 dx}_{\mathtt{A}} + \alpha^r \sum_{j=1}^{2} \underbrace{\int_\Omega \|D\sigma_j\|^2 dx}_{\mathtt{B}}, \tag{2.3}$$

where α^r is a positive constant. The interpretation of this functional is the following: we would like the OFC to be zero (term A) and the gradient magnitude to be minimum (term B). Notice that term B is the classical Tikhonov–Arsenin [48] relaxation known to smooth isotropically. With this method, we obtain a smooth optical flow, and flow discontinuities are lost.

2.3. Setting the problem. The purpose of this work is to propose a model able to cope with the discontinuities of the optical flow. Starting from (2.3), we propose to minimize the energy

$$E(\sigma) = \underbrace{\int_\Omega |(\sigma \cdot Du) + u_t|}_{\mathtt{A}} + \alpha^r \underbrace{\sum_{j=1}^{2} \int_\Omega \phi(D\sigma_j)}_{\mathtt{B}} + \alpha^h \underbrace{\int_\Omega c(x)\|\sigma\|^2 dx}_{\mathtt{C}}, \tag{2.4}$$

where α^r, α^h are positive constants, $\phi(\cdot)$ and $c(\cdot)$ to be determined. We refer the interested reader to [35, 6] for the detailed construction of this model. Let us describe briefly the main differences:

(i) Term `A` is comparable to term `A` in (2.3). Here we choose the L^1 norm which must be interpreted in term of measures. As we will see in what follows, since the data u belongs a priori to $BV(\Omega)$, we cannot use the L^2 norm as done in (2.3).

(ii) Term `B` is again a regularization term. The functions $\phi(\cdot)$ have been chosen so that we can preserve discontinuities. The key idea is to forbid smoothing across discontinuities. Such ideas initially were proposed in the image restoration background [45, 20, 5, 7], and many functions have been proposed. Typically, admissible functions are convex functions with linear growth at infinity. For instance, we will choose the minimal hypersurface function

$$\phi(s) = \sqrt{s^2 + 1}.$$

We mention that term `B` will be interpreted as convex functions of measures.

(iii) Finally, term `C` permits us to handle the homogeneous regions. Typically, $c(x)$ is high for low spatial gradients of u (hence penalizing velocities in poor information zones) and low for high spatial gradients of u (no intervention).

3. General recalls. In this section we recall main notations and definitions. We refer to [1, 22, 25, 23, 53] for the complete theory.

Let Ω be a bounded open set in R^N with Lipschitz-regular boundary $\partial\Omega$. We denote by $\mathcal{L}^N$ or dx the N-dimensional Lebesgue measure in R^N and by $\mathcal{H}^\alpha$ the α-dimensional Hausdorff measure. We also set $|E| = \mathcal{L}^N(E)$, the Lebesgue measure of a measurable set $E \subset R^N$. $\mathcal{B}(\Omega)$ denotes the family of the Borel subsets of Ω. We will denote the strong, the weak, and weak$\star$ convergences in a space $V(\Omega)$ by $\xrightarrow[V(\Omega)]{}$, $\xrightarrow[V(\Omega)]{}$, and $\xrightarrow[V(\Omega)]{\star}$, respectively. Spaces of vector-valued functions will be denoted by bold characters.

Working with images requires that the functions we consider can be discontinuous along curves. This is impossible with classical Sobolev spaces such as $W^{1,1}(\Omega)$. This is why we need to use the space of bounded variations ($BV(\Omega)$) defined by

$$BV(\Omega) = \left\{ u \in L^1(\Omega); \sup \int_\Omega u \mathrm{div}(\varphi) dx < \infty : \varphi \in \mathcal{C}_0^1(\Omega)^N, |\varphi|_\infty \leq 1 \right\},$$

where $\mathcal{C}_0^1(\Omega)$ is the set of differentiable functions with compact support in Ω. We will note

$$|Du|(\Omega) = \sup \left\{ \int_\Omega u \mathrm{div}(\varphi) dx : \varphi \in \mathcal{C}_0^1(\Omega)^2, |\varphi|_\infty \leq 1 \right\}.$$

If $u \in BV(\Omega)$ and Du is the gradient in the sense of distributions, then Du is a vector-valued Radon measure and $|Du|(\Omega)$ is the total variation of Du on Ω. The set of Radon measure is noted $\mathcal{M}(\Omega)$

The product topology of the strong topology of $L^1(\Omega)$ for u and of the weak$\star$ topology of measures for Du will be called the weak$\star$ topology of BV and will be denoted by $BV - w\star$.

$$u^n \xrightarrow[BV-w\star]{} u \iff \begin{cases} u^n \xrightarrow[L^1(\Omega)]{} u, \\ Du^n \xrightarrow[\mathcal{M}(\Omega)]{\star} Du. \end{cases} \tag{3.1}$$

We recall that every bounded sequence in $BV(\Omega)$ admits a subsequence converging in $BV - w\star$.

We define the approximate upper limit $u^+(x)$ and the approximate lower limit $u^-(x)$ by

$$u^+(x) = \inf\left\{t \in [-\infty, +\infty] : \lim_{\rho \to 0^+} \frac{\mathcal{L}^N\left(\{u > t\} \cap B_\rho(x)\right)}{\rho^N} = 0\right\},$$
$$u^-(x) = \sup\left\{t \in [-\infty, +\infty] : \lim_{\rho \to 0^+} \frac{\mathcal{L}^N\left(\{u < t\} \cap B_\rho(x)\right)}{\rho^N} = 0\right\},$$

where $B_\rho(x)$ is the ball of center x and radius ρ. We denote by S_u the jump set, that is, the complement up to a set of $\mathcal{H}^{N-1}$ measures zero of the set of Lebesgue points; i.e., the set of points x where $u^+(x)$ is different $u^-(x)$, namely,

$$S_u = \{x \in \Omega / u^-(x) < u^+(x)\}.$$

After choosing a normal $n_u(x)$ $(x \in S_u)$ pointing toward the largest value of u, we recall the following decompositions (see [2] for more details):

$$Du = \nabla u \cdot \mathcal{L}_N + C_u + (u^+ - u^-) n_u \cdot \mathcal{H}^{N-1}_{|S_u}, \tag{3.2}$$

$$|Du|(\Omega) = \int_\Omega \|\nabla u\| dx + \int_{\Omega \setminus S_u} |C_u| + \int_{S_u} (u^+ - u^-) d\mathcal{H}^{N-1}, \tag{3.3}$$

where ∇u is the density of the absolutely continuous part of Du with respect to the Lebesgue measure, C_u is the Cantor part, and $\mathcal{H}^{N-1}$ is the Hausdorff measure of dimension $N-1$.

We then recall the definition of a convex function of measures. We refer to the works of Goffman–Serrin [26] and Demengel–Temam [19] for more details. Let $\phi(\cdot)$ be convex and finite on R with linear growth at infinity. Let ϕ^∞ be the asymptote (or recession) function defined by $\phi^\infty(z) = \lim_{s \to \infty} \frac{\phi(sz)}{s} \in [0; +\infty)$. Then, for $u \in BV(\Omega)$, using classical notations, we define

$$\int_\Omega \phi(Du) = \int_\Omega \phi(\|\nabla u\|) dx + \phi^\infty(1) \int_{S_u} (u^+ - u^-) d\mathcal{H}^{N-1} + \phi^\infty(1) \int_{\Omega \setminus S_u} |C_u|. \tag{3.4}$$

Finally, we mention that this function is lower semicontinuous for the $(BV - w\star)$-topology.

4. The integral representation of the optical flow problem.

4.1. The precise formulation. This section is devoted to the mathematical study of the optical flow model proposed in section 2.3. Let us recall it. Without loss of generality, we will assume that $\alpha^r = \alpha^h = 1$. For $u \in BV(R \times \Omega)$, the problem is to find σ minimizing the energy

$$E(\sigma) = \int_\Omega |(\sigma \cdot Du) + u_t| + \sum_{j=1}^{N} \int_\Omega \phi(D\sigma_j) + \int_\Omega c(x)\|\sigma\|^2 dx, \tag{4.1}$$

where the infimum is taken over the space $\mathbf{BV}(\Omega)$. The function $\phi(\cdot)$ verifies

$$\phi : R \to R^+ \text{ is an even and convex function, nondecreasing on } R^+. \tag{4.2}$$

There exist constants $d > 0$ and $b \geq 0$ such that

$$dx - b \leq \phi(x) \leq dx + b \text{ for all } x \in R, \tag{4.3}$$

$$\phi^\infty(1) = 1, \tag{4.4}$$

$$\phi^*(x^*) \leq k, \qquad \text{for all } x^* \in dom(\phi^*) \qquad (k \text{ constant}), \tag{4.5}$$

where ϕ^* is the Legendre–Fenchel conjugate function of ϕ. Note that hypotheses (4.3) and (4.5) permit the assertion that (see [43])

$$\phi(x) \geq \phi^\infty(x) - k. \tag{4.6}$$

Finally, we assume that the function $c(\cdot)$ verifies the following assumptions:

$$c \in \mathcal{C}^\infty(\Omega), \tag{4.7}$$

$$\text{there exists a constant } m_c > 0 \text{ such that } c(x) \in [m_c, 1] \text{ for all } x \text{ in } \Omega. \tag{4.8}$$

4.2. The duality pairing $(\sigma \cdot Du)$: An extended integral representation. This part is devoted to a better understanding of the functional to be minimized (4.1) and especially the product $(\sigma \cdot Du)$. In fact, what can we say about the product of an integrable function and a measure? This question has been treated for special cases, with suitable hypotheses on σ and u (see [3, 47, 12]). For example, Anzellotti [3] supposes

$$\sigma \in \mathbf{X}(\Omega) \cap \mathcal{C}^0(\Omega; R^N) \text{ and } u(t_0, \cdot) \in BV(\Omega),$$
$$\sigma \in \mathbf{X}(\Omega) \text{ and } u(t_0, \cdot) \in W^{1,1}(\Omega),$$

where $\mathbf{X}(\Omega) = \{\sigma \in \mathbf{L}^\infty(\Omega);\ \mathrm{div}(\sigma) \in L^N(\Omega)\}$. Our aim is to extend his results for a more general class of product $(\sigma \cdot Du)$. We suppose that

$$\sigma \in \mathbf{BV}(\Omega) \cap \mathbf{X}(\Omega), \tag{4.9}$$

$$u(t_0, \cdot) \in SBV(\Omega) \cap L^\infty(\Omega), \tag{4.10}$$

where $SBV(\Omega)$ is the space of special bounded variations (the Cantor part of Du is zero).

Remark. The hypothesis (4.10) is quite general. We mention to the interested reader a more applied work where we assumed only that the data u is Lipschitz [6]. In that case, there is no problem to define the L^1 norm of the optical flow constraint,

and we proved the existence and uniqueness of the minimization problem posed on $\mathbf{BV}(\Omega)$. We also proposed a convergent algorithm to approximate the solution (using Γ-convergence arguments) and showed some numerical results on synthetic and real sequences.

The space $\mathbf{X}(\Omega)$ is a Banach space endowed with the norm

$$\|\sigma\|_{\mathbf{X}(\Omega)} = \|\sigma\|_{\mathbf{L}^\infty(\Omega)} + \|\mathrm{div}(\sigma)\|_{L^N(\Omega)},$$

and we can define a weak$\star$ topology on $\mathbf{X}(\Omega)$ by

$$\sigma^n \xrightarrow[\mathbf{X}(\Omega)]{} \sigma \quad \Longleftrightarrow \quad \begin{cases} \sigma^n \xrightarrow[\mathbf{L}^\infty(\Omega)]{\star} \sigma, \\ \mathrm{div}(\sigma^n) \xrightarrow[L^N(\Omega)]{} \mathrm{div}(\sigma). \end{cases}$$

To make sense of the pairing $(\sigma \cdot Du)$, our first thought is to define it by duality:

$$\int_\Omega \varphi(\sigma \cdot Du) = -\int_\Omega u\varphi \mathrm{div}(\sigma)dx - \int_\Omega u\sigma\nabla\varphi dx \quad \text{for all} \quad \varphi \ \text{ in } \ \mathcal{C}_c^1(\Omega). \tag{4.11}$$

Note that with hypotheses (4.9) and (4.10), the right-hand side of (4.11) is completely defined. We can prove [3, 12] that $(\sigma \cdot Du)$ is a bounded measure, absolutely continuous with respect to $|Du|$.

Our aim is to find an integral representation of that measure.

To this end, we need to introduce the precise representation of σ, noted $\tilde{\sigma}$ and defined by

$$\tilde{\sigma}(x) = \lim_{r\to 0} \frac{1}{\mathcal{L}^N(B(x,r))} \int_{B(x,r)} \sigma(y)dy. \tag{4.12}$$

If σ is simply in $\mathbf{L^1}(\Omega)$, the right-hand side limit exists $\mathcal{L}^N$ almost everywhere (a.e.) and is equal to $\sigma(x)$. However, if σ is also in $\mathbf{BV}(\Omega)$, we can explicitely write the limit $\mathcal{H}^{N-1}$ a.e. using σ^+, σ^-. We have [52, 22]

$$\tilde{\sigma}(x) = \frac{\sigma^+(x) + \sigma^-(x)}{2}, \qquad \mathcal{H}^{N-1} \text{ a.e. on } S_\sigma. \tag{4.13}$$

Another interesting property of $\tilde{\sigma}$ is that we have the approximation result

$$\tilde{\sigma}(x) = \lim_{\varepsilon\to 0} \eta_\varepsilon \star \sigma(x), \qquad \mathcal{H}^{N-1} \text{ a.e.}, \tag{4.14}$$

where (η_ε) are the usual mollifiers: $\eta_\varepsilon \in \mathcal{C}_c^\infty(R^N), spt(\eta_\varepsilon) \subset B(0,\varepsilon), 0 \le \eta_\varepsilon \le 1$, $\int_{R^N} \eta_\varepsilon(x)dx = 1$. The function $\tilde{\sigma}$ is called the *precise representation* of σ since it permits us in some way to define $\sigma - \mathcal{H}^{N-1}$ a.e. Remark that $\tilde{\sigma}$ and σ are in fact the same elements in $\mathbf{BV}(\Omega)$ (they belong to the same equivalent class of $\mathcal{L}^N$a.e. equal functions) so that their distributional derivatives are the same.

From now on, we will consider that $N = 2$. For two measures μ and ν in $\mathcal{M}(\Omega)$, we will denote by $\frac{d\mu}{d\nu}$ the Radon–Nikodym derivative of μ with respect to ν (see [22] for more details).

The main result of this section is the following proposition.

PROPOSITION 4.1. *If $\sigma \in \mathbf{X}(\Omega) \cap \mathbf{BV}(\Omega)$ and $u(t_0, \cdot) \in SBV(\Omega) \cap L^\infty(\Omega)$, then we have*

$$\int_B (\sigma \cdot Du) = \int_B \tilde{\sigma}(x) \cdot \frac{dDu}{d|Du|}(x)|Du| \quad \text{for all Borel set } B \text{ in } \Omega. \tag{4.15}$$

Moreover, if $u \in SBV(\Omega)$, we obtain

$$\int_\Omega (\sigma \cdot Du) = \int_\Omega \sigma \cdot \nabla u dx + \int_{S_u} \tilde{\sigma} \cdot n_u (u^+ - u^-) d\mathcal{H}^{N-1}. \tag{4.16}$$

Before proving this result, we mention a convergence result which can be demonstrated using arguments from [3].

LEMMA 4.2. *Let $\sigma_\varepsilon = \eta_\varepsilon \star \sigma(x)$. If $\sigma \in \mathbf{BV}(\Omega) \cap \mathbf{X}(\Omega)$, then we have*

$$\sigma_\varepsilon \xrightarrow[\mathbf{L}^\infty(A)]{\star} \sigma, \tag{4.17}$$

$$\operatorname{div}(\sigma_\varepsilon) \xrightarrow[L^p(A)]{} \operatorname{div}(\sigma), \qquad p < \infty, \tag{4.18}$$

for all open sets $A \subset \Omega$. Moreover, for all $u \in BV_{loc}(\Omega) \cap L^\infty(\Omega)$, one has

$$(\sigma_\varepsilon \cdot Du) \xrightarrow[\mathcal{M}(\Omega)]{} (\sigma \cdot Du). \tag{4.19}$$

Proof of Proposition 4.1. If we denote $\sigma_\varepsilon = \eta_\varepsilon \star \sigma$, then for all $\varphi \in \mathcal{C}_c^1(\Omega)$, we have (see Lemma 4.2)

$$\langle (\tilde{\sigma} \cdot Du), \varphi \rangle = \lim_{\varepsilon \to 0} \langle (\sigma_\varepsilon \cdot Du), \varphi \rangle. \tag{4.20}$$

As $Du \ll |Du|$, by the Radon–Nikodym theorem, there exists a function $h \in L^1_{|Du|}(\Omega)$, $|h(x)| = 1$, such that $Du = h|Du|$. Thus, since $\sigma_\varepsilon \in L^\infty_{|Du|}(\Omega)$ ($\sigma_\varepsilon \in \mathcal{C}^\infty(\Omega)$) we can write

$$\langle (\sigma_\varepsilon \cdot Du), \varphi \rangle = \int_\Omega \varphi \sigma_\varepsilon \cdot h|Du|. \tag{4.21}$$

Equations (4.20) and (4.21) imply that

$$\langle (\sigma \cdot Du), \varphi \rangle \;=\; \lim_{\varepsilon \to 0} \int_\Omega \varphi \sigma_\varepsilon \cdot h|Du|. \tag{4.22}$$

What remains is to show the permutation between the limit and the integral in (4.22). To do this, we use the Lebesgue dominated convergence theorem. Classically, two requirements are necessary:

• The pointwise convergence of $\varphi(x)\sigma_\varepsilon(x) \cdot h(x)$ to $\varphi(x)\tilde{\sigma}(x) \cdot h(x)$. It comes from (4.14). Notice that the pointwise convergence is true $\mathcal{H}^{N-1}$ a.e. and consequently $|Du|$ a.e.

• Find a function which dominates the sequence. In fact, since Ω is bounded, it is sufficient to prove that the L^∞ norm of $\varphi(x)\sigma_\varepsilon(x) \cdot h(x)$ is bounded uniformly by a constant. Since φ is in $\mathcal{C}_c^1(\Omega)$ and $|h(x)| = 1$, it is enough to show that there exists a constant C such that $\|\sigma_\varepsilon\|_{\mathbf{L}^\infty{}_{Du}(\Omega)} \le C$. In fact we have

$$\|\sigma_\varepsilon\|_{\mathbf{L}^\infty_{\mathbf{Du}}(\Omega)}, \le \underset{R^N}{\sup} \; |\sigma_\varepsilon| \;\underset{[*]}{\le} \|\sigma\|_{\mathbf{L}^\infty_{\mathcal{L}^\mathbf{N}}(\Omega)} \le C,$$

where inequality [∗] is shown in [3]. Consequently, we can apply the Lebesgue dominated convergence theorem. This permits us to pass to the limit in (4.22) as $\varepsilon \to 0$, and we get (4.15). It is then an easy task to get (4.16) from (4.15) using the decomposition (3.2). □

4.3. Application to the optical flow problem. Now that we have found an expression of the product $(\sigma \cdot Du)$, we will give in the next proposition the integral representation of the functional E, which will be used in what follows.

We will assume that

$$u \in SBV(R \times \Omega) \cap L^{\infty}(R \times \Omega). \tag{4.23}$$

There exists $h_1 \in L^1(\Omega)$ and $h_2 \in L^1_{\mathcal{H}^1}(S_u)$ such that

$$u_t = h_1 \mathcal{L}^2 + h_2 \mathcal{H}^1|_{S_u}. \tag{4.24}$$

Notice that the assumption means that the measure u_t is absolutely continuous with respect to $|Du|$. This is physically correct, since when there is no texture (no gradient), no intensity variation should be observed.

PROPOSITION 4.3. *We assume that $N = 2$. Let u verifying hypotheses* (4.23)–(4.24). *Then the function E defined on $\mathbf{X} \cap \mathbf{BV}(\Omega)$ by*

$$E(\sigma) = \int_{\Omega} |(\sigma \cdot Du) + u_t| + \sum_{j=1}^{2} \int_{\Omega} \phi(D\sigma_j) + \int_{\Omega} c(x) \|\sigma\|^2 dx \tag{4.25}$$

with hypotheses (4.2)–(4.3), (4.4)–(4.6), (4.7)–(4.8)*can be rewritten as*

$$\begin{aligned} E(\sigma) = & \int_{\Omega} |\sigma \cdot \nabla u + h_1| dx + \int_{S_u} |\tilde{\sigma} \cdot n_u (u^+ - u^-) + h_2| d\mathcal{H}^1 \\ & + \sum_{j=1}^{2} \int_{\Omega} \phi(D\sigma_j) + \int_{\Omega} c(x) \|\sigma\|^2 dx. \end{aligned} \tag{4.26}$$

Proof. Thanks to [46, Theorem 6.13], we know that if ν is a positive measure on $\mathcal{M}(\Omega)$, $g \in L^1_{\nu}$, and λ is the measure defined by

$$\lambda(E) = \int_E g d\nu,$$

then we have

$$|\lambda|(E) = \int_E |g| d\nu.$$

Moreover, using the decomposition of the measure u_t and the result (4.16), we have

$$\int_{\Omega} (\sigma \cdot Du) + u_t = \int_{\Omega} (\sigma \cdot \nabla u + h_1) dx + \int_{S_u \cap \Omega} (\tilde{\sigma} \cdot n_u (u^+ - u^-) + h_2) d\mathcal{H}^1.$$

Using the fact that dx and $d\mathcal{H}^1$ are mutually singular and applying the above theorem, we can conclude the proof. □

COMMENTARY ABOUT PROPOSITION 4.3. The interesting point in the integral representation (4.26) is that we no longer need the divergence of σ to be integrable. Consequently, (4.26) can be viewed as an extension of E defined a priori for $\sigma \in \mathbf{BV}(\Omega)$. The next section is devoted to the theoretical study of that extension.

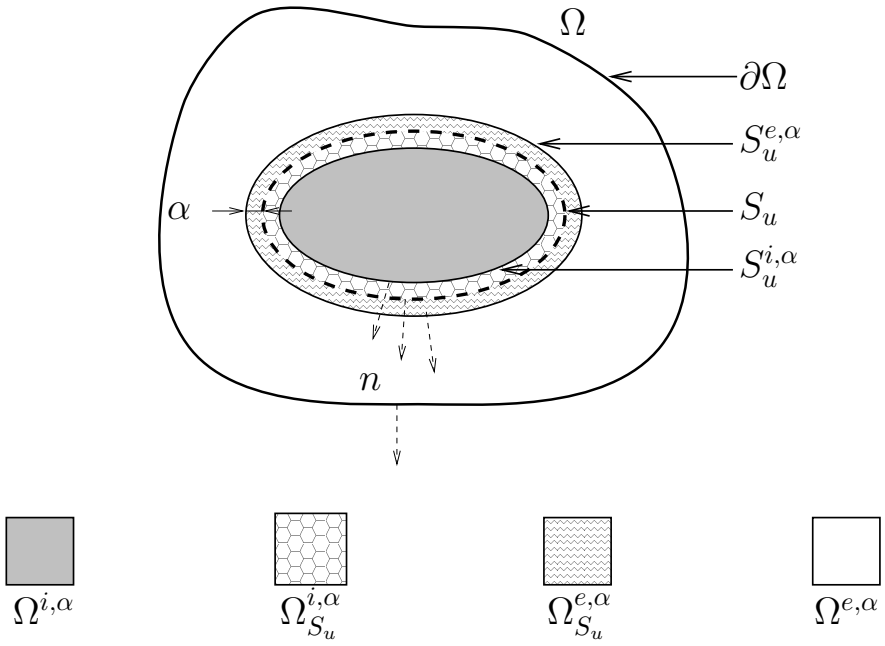


FIG. 5.1. *Notation of the simplified problem. Notice that the normals are all oriented toward the exterior.*

5. The relaxed problem. After introducing notation and assumptions in section 5.1, we show in section 5.2 that the functional that we are considering is not lower semicontinuous for the $BV - w\star$ topology. As a consequence, the existence of a solution cannot be shown for the initial problem. We then search for the relaxed functional for a suitable topology in section 5.4 after proving some preliminary results in section 5.3.

5.1. Notation and assumptions. To simplify proofs and notation, we will assume in this section that $N = 2$ and that S_u is a single $\mathcal{C}^1$ curve as shown in Figure 5.1, where the main notation is introduced. Notice that the parameter α corresponds to the distance between S_u and $S_u^{e,\alpha}$ (or $S_u^{i,\alpha}$). We will also use the superscript i (respectively, e) to mention that we are considering the restriction of the function to $\Omega^i \equiv \Omega^{i,0}$ (respectively, $\Omega^e \equiv \Omega^{e,0}$). The Hausdorff measure of dimension 1 is denoted by ds.

Using this notation we rewrite the integral on S_u of (4.26), which is

$$\int_{S_u} |\tilde{\sigma} \cdot n_u(u^+ - u^-) + h_2| ds. \tag{5.1}$$

Let $b = \pm 1$ be the function such that $n_u = bn$, where n is the normal oriented toward the exterior (Voir Figure 5.1). Let $\tilde{h}_2$ be the function defined by $\tilde{h}_2 = bh_2$. It is then easy to check that (5.1) may be rewritten as

$$\int_{S_u} \left| \frac{\sigma^i + \sigma^e}{2} \cdot n(u^+ - u^-) + \tilde{h}_2 \right| ds. \tag{5.2}$$

So, changing h_2 in $\tilde{h}_2$ allows us to have a normal independent of u. We will use this

expression, which is easier to handle. To simplify notation, we will omit the tilde superscript for h_2.

5.2. Statement of the problem. Let us first recall precisely the problem that we are going to study. Let $h_1 \in L^1(\Omega)$ and $h_2 \in L^1_{\mathcal{H}^1}(S_u)$. Let $\phi(\cdot)$ be a function verifying (4.2)–(4.3), (4.4)–(4.6), and c satisfying (4.7)–(4.8). Let E be the functional defined over $\mathbf{BV}(\Omega)$ by

$$\begin{aligned} E(\sigma) = & \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \int_{S_u} |\tilde{\sigma} \cdot n(u^+ - u^-) + h_2| ds \\ & + \sum_{j=1}^{2} \int_\Omega \phi(D\sigma_j) + \int_\Omega c(x) \|\sigma\|^2 dx. \end{aligned} \tag{5.3}$$

We remark that it is well defined on $\mathbf{BV}(\Omega)$ thanks to the embedding of $\mathbf{BV}(\Omega)$ into $\mathbf{L^2}(\Omega)$ ($N = 2$) (see, for instance, [25]). Our aim is to study the existence of a solution to the minimization problem

$$\inf_{\sigma \in \mathbf{BV}(\Omega)} E(\sigma) \ . \tag{5.4}$$

Following the direct method of the calculus of variations, let (σ^n) be a minimizing sequence of (5.3). Thanks to hypotheses on functions $\phi(\cdot)$ and $c(\cdot)$, we can obtain a uniform majoration in $\mathbf{BV}(\Omega)$ and in $\mathbf{L^2}(\Omega)$, so we can extract a subsequence converging to some σ for the topology $BV - w\star$ and L^2-weak. The question is, Can we deduce an existence result for (5.4)? To answer this question, let us split the functional E in two parts, namely, P and L, defined by

$$P(\sigma) = \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \sum_{j=1}^{2} \int_\Omega \phi(D\sigma_j) + \int_\Omega c(x) \|\sigma\|^2 dx, \tag{5.5}$$

$$L(\sigma) = \int_{S_u} \left| \frac{\sigma^i + \sigma^e}{2} \cdot n(u^+ - u^-) + h_2 \right| ds. \tag{5.6}$$

It is easy to show that we have

$$\lim_n P(\sigma^n) \geq P(\sigma),$$

but we cannot say anything about the term L. The reason is that the functional L is defined through traces and the trace application is not continuous for the weak$\star$ topology of $\mathbf{BV}(\Omega)$. Consequently, the functional E is not lower semicontinuous for the $BV - w\star$ topology. In such a situation, the idea is to study the relaxed functional.

We recall that for a functional F defined over a topological metrizable space X, the relaxed functional, noted $R(F)$, verifies that

$$\text{for all } u \in X, \forall u^n \to u. \liminf_n F(u^n) \geq R(F)(u), \tag{5.7}$$

$$\text{for all } u \in X, \exists u^n \to u. \limsup_n F(u^n) \leq R(F)(u). \tag{5.8}$$

$R(F)$ is in fact the higher lower semicontinuous functional less than or equal to F. We refer the interested reader to [36, 16] for a complete overview of the relaxation properties and consequences.

5.3. Preliminary results. As is usual when we have this kind of problem, we need to introduce additional variables and notation. Let us define the functionals $\tilde{L}$ and E_1 by

$$\tilde{L} : \mathbf{M}(S_u) \times \mathbf{M}(S_u) \to R, \qquad \tilde{L}(\mu^i, \mu^e) = \int_{S_u} d|\nu|, \tag{5.9}$$

where

$$\nu = \frac{\mu^i + \mu^e}{2} \cdot n(u^+ - u^-) + h_2 ds \tag{5.10}$$

and

$$E_1 : \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u) \to R, \qquad E_1(\sigma, \mu^i, \mu^e) = \begin{cases} P(\sigma) + \tilde{L}(\mu^i, \mu^e) & \text{if } \mu^i = \sigma^i ds \text{ and } \mu^e = \sigma^e ds, \\ +\infty & \text{otherwise.} \end{cases} \tag{5.11}$$

It is straightforward to see that

$$\inf_{\sigma \in BV(\Omega)} E(\sigma) = \inf_{(\sigma, \mu^i, \mu^e) \in \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)} E_1(\sigma, \mu^i, \mu^e). \tag{5.12}$$

The functionals (5.3) and (5.11) are not weakly lower semicontinuous, so it is natural to search for the relaxed functionals of E and E_1, noted $R(E)$ and $R(E_1)$, for a suitable topology.

Thanks to classical results [36, 16], we have, using (5.12),

$$\inf_{\sigma \in BV(\Omega)} E(\sigma) = \inf_{\sigma \in BV(\Omega)} R(E)(\sigma) = \inf_{(\sigma, \mu^i, \mu^e) \in \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)} R(E_1)(\sigma, \mu^i, \mu^e).$$

Moreover, since the relaxed functionals are lower semicontinuous, existence results can be proved. Our aim is then to compute these relaxed functionals, which is the main result of section 5. To this end, we will use the definitions (5.7) and (5.8). Difficulties are twofold:

- We must guess the expression of the functional which is a priori unknown. This will be done using the property (5.7) with some care.
- To check that the guess is really the relaxed functional, we need to verify (5.8). The main difficulty is that we must find the sequence (u^n) converging to a given u. However, we will see how we can avoid this difficulty.

We mention that the notion of relaxation is classical in many problems occurring in the calculus of variations: phase transition, fracture mechanics, and plasticity, to name a few. For recent advances and bibliography, we refer to [10].

The specificity of this work is that the surface energy is defined over a fixed set independent of the unknown σ. Moreover, we give an explicit representation of the

relaxed energy. We are going to establish that the functional $\overline{E}_1$ defined by

(5.13)

$$\overline{E}_1 : \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u) \to R,$$
$$\overline{E}_1(\sigma, \mu^i, \mu^e) = \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \sum_{j=1}^{2} \int_{\Omega^i \cup \Omega^e} \phi(D\sigma_j) + \int_\Omega c(x) \|\sigma\|^2 + \int_{S_u} d|\nu| + \int_{S_u} \|\mu^i - \mu^e\|_1 + \int_{S_u} \|\mu^i - \sigma^i ds\|_1 + \|\mu^e - \sigma^e ds\|_1,$$

where $\|\eta\|_1 = |\eta_1| + |\eta_2|$ and the measure ν is defined by (5.10), is in fact the relaxed functional of E_1 for the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak. We are also going to prove that the functional defined by

(5.14)

$$\overline{E} : \mathbf{BV}(\Omega) \to R,$$
$$\overline{E}(\sigma) = \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \sum_{j=1}^{2} \int_{\Omega^i \cup \Omega^e} \phi(D\sigma_j) + \int_\Omega c(x) \|\sigma\|^2 + \int_{S_u} \beta(x, \sigma^i, \sigma^e),$$

where

(5.15)

$$\beta(x, \lambda, \theta) = \mathrm{Inf} \left\{ |\lambda - s| + |\theta - t| + |s - t| + \left| \frac{s+t}{2} \cdot n(x)(u^+ - u^-) + h_2(x) \right| \ : \ (s,t) \in R^N \times R^N \right\},$$

is the relaxed functional of E. The expression of $\overline{E}$ will be deduced from $\overline{E_1}$.

Before finding (5.13) and (5.14), we first need to prove some preliminary results. The general idea is that, for technical reasons, we need to work with functions defined on more regular spaces. This is why we introduce the functionals E_2 and $\overline{E}_2$ defined by

(5.16)
$$E_2 : \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u) \to R,$$
$$E_2(\sigma, \mu^i, \mu^e) = \begin{cases} E_1(\sigma, \mu^i, \mu^e) & \text{if } \ \sigma \in \mathbf{W^{1,1}}(\Omega^i \cup \Omega^e), \\ +\infty & \text{otherwise} \end{cases}$$

and

(5.17)
$$\overline{E}_2 : \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u) \to R,$$
$$\overline{E}_2(\sigma, \mu^i, \mu^e) = \begin{cases} \overline{E}_1(\sigma, \mu^i, \mu^e) & \text{if } (\sigma, \mu^i, \mu^e) \in \mathbf{W^{1,1}}(\Omega^i \cup \Omega^e) \times \mathbf{L^1}(S_u) \times \mathbf{L^1}(S_u), \\ +\infty & \text{otherwise.} \end{cases}$$

The justification of considering $E_2, \overline{E}_2$ instead of $E_1, \overline{E}_1$ is given by Lemmas 5.2 and 5.3, where we prove that E_j and $\overline{E}_j (j = 1, 2)$ have the same relaxed functional for the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak. As this is equivalent to saying that they have the same dual functional [13], we will use (5.16) and (5.17) to

compute the dual functionals (Lemmas A.1, A.2) and to establish the main relaxation result.

Let us present a version of the slicing lemma of De Giorgi that will be useful in what follows.

THEOREM 5.1. *Let $\phi(\cdot)$ be a function verifying hypotheses* (4.2)–(4.3). *Let $u \in BV(\Omega) \cap L^2(\Omega)$. Then, for every open set $A \subset \Omega$ with Lipschitz boundary, we can find a sequence $u^n \in W^{1,1}(\Omega)$ such that*

$$u^n \xrightarrow[L^2(\Omega)]{} u, \tag{5.18}$$

$$u^n = u \quad on \quad \partial A, \tag{5.19}$$

$$\lim_{n\to\infty} \int_A \phi(\|\nabla u^n\|)dx = \int_A \phi(Du). \tag{5.20}$$

Notice that this theorem permits us to fix the trace at the boundaries.

Proof. The proof of this theorem is a consequence of Lemma 2.6 proposed in [11], which can be modified to obtain the strong convergence in L^2. □

LEMMA 5.2. *Let E_1 and E_2 be defined, respectively, by* (5.11) *and* (5.16). *Then E_1 and E_2 have the same lower semicontinuous relaxed functions for the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak.*

Proof. The proof contains two steps.

Step 1. Since we have $E_1 \leq E_2$, we deduce that

$$R(E_1) \leq R(E_2). \tag{5.21}$$

Step 2. The reverse inequality will be proven using an approximation argument. Let $(\sigma, \mu^i, \mu^e) \in \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)$ such that E_1 be finite. Notice that this forces the measures μ^i, μ^e to be the traces of σ. We search for a sequence $(\sigma^n, \mu^{i^n}, \mu^{e^n}) \in \mathbf{W^{1,1}}(\Omega^i \cup \Omega^e) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)$ converging to (σ, μ^i, μ^e) for the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak such that

$$\lim_{n\to\infty} E_2(\sigma^n, \mu^{i^n}, \mu^{e^n}) = E_1(\sigma, \mu^i, \mu^e). \tag{5.22}$$

If we can find such a sequence, then the proof is complete since equality (5.22) means that $E_1 \geq R(E_2)$. However, since $R(E_1)$ is the greatest lower semicontinuous function less than or equal to E_1, we deduce that

$$R(E_1) \geq R(E_2). \tag{5.23}$$

Inequalities (5.21) and (5.23) conclude the proof. The difficulty lies in finding such a sequence. The idea is to apply Theorem 5.1 in Ω^i and Ω^e, separately. In Ω^i, we obtain the existence of a sequence (σ^{i^n}) such that

$$\sigma^{i^n} \xrightarrow[L^2(\Omega^i)]{} \sigma,$$

$$\sigma^{i^n}|_{S_u} = \sigma^i|_{S_u},$$

$$\lim_{n\to\infty} \int_{\Omega^i} \phi(\|\nabla \sigma^{i^n}\|)dx = \int_{\Omega^i} \phi(D\sigma). \tag{5.24}$$

We proceed as we had done in Ω^e and define the sequence (σ^n)-$\mathcal{L}^2$ a.e. on Ω by

$$\sigma^n(x) = \begin{cases} \sigma^{i^n} & \text{if } x \in \Omega^i, \\ \sigma^{e^n} & \text{if } x \in \Omega^e. \end{cases}$$

It is easy to check that σ^n belongs to $\mathbf{W^{1,1}}(\Omega^i \cup \Omega^e)$. Using that sequence, we define the sequence of measures ${\mu^i}^n$ and ${\mu^e}^n$ defined on S_u by

$$ {\mu^i}^n = {\sigma^i}^n ds, \tag{5.25} $$

$$ {\mu^e}^n = {\sigma^e}^n ds. \tag{5.26} $$

Notice that since we have fixed the traces of σ^n on both sides of S_u, the sequences defined by (5.25)–(5.26) are in fact constant. So we have

$$ E_2(\sigma^n, {\mu^i}^n, {\mu^e}^n) = \underbrace{\int_\Omega |\sigma^n \cdot \nabla u + h_1| dx + \int_\Omega c(x) \|\sigma^n\|^2 dx}_{\text{continuous for the } L^2 \text{ strong topology}} + \sum_{j=1}^{2} \int_\Omega \phi(D\sigma_j^n) + \underbrace{\int_{S_u} d|\nu|}_{\text{constant}} $$

and, moreover,

$$ \begin{aligned} \int_\Omega \phi(D\sigma_j) &= \int_{\Omega^i} \phi(D\sigma_j) + \int_{\Omega^e} \phi(D\sigma_j) + \int_{S_u} |\sigma_j^i - \sigma_j^e| ds \\ &= \lim_{n\to\infty} \int_{\Omega^i} \phi(\|\nabla {\sigma_j^i}^n\|) dx + \lim_{n\to\infty} \int_{\Omega^e} \phi(\|\nabla {\sigma_j^e}^n\|) dx + \int_{S_u} |\sigma_j^i - \sigma_j^e| ds \\ &= \lim_{n\to\infty} \int_\Omega \phi(\|\nabla \sigma_j^n\|) dx. \end{aligned} $$

Thus condition (5.22) is satisfied, and this concludes the proof. ∎

LEMMA 5.3. *Let $\overline{E}_1$ and $\overline{E}_2$ be defined by* (5.13) *and* (5.17)*, respectively. Then $\overline{E}_1$ and $\overline{E}_2$ have the same lower semicontinuous relaxed functionals for the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak.*

Proof. This proof is inspired by the proof of Lemma 5.2. The first step is analogous, and the only difficulty is to find, for a given $(\sigma, \mu^i, \mu^e) \in \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)$, a sequence $(\sigma^n, {\theta^i}^n, {\theta^e}^n) \in \mathbf{W^{1,1}}(\Omega^i \cup \Omega^e) \times \mathbf{L^1}(S_u) \times \mathbf{L^1}(S_u)$ such that

$$ \lim_{n\to\infty} \overline{E}_2(\sigma^n, {\theta^i}^n, {\theta^e}^n) = \overline{E}_1(\sigma, \mu^i, \mu^e). \tag{5.27} $$

Construction of the sequence σ^n uses the same arguments as in Lemma 5.2, that is, the use of Theorem 5.1 on Ω^i and Ω^e. We recall that the traces of σ^n on both sides of S_u are constant. The construction of the sequence approximating μ^i, μ^e is based on a result of Bouchitté–Valadier [14]. We recall that the part depending on the measures μ^i, μ^e in $\overline{E}_1$ is

$$ H(x, \mu^i, \mu^e, \sigma^i, \sigma^e) = \int_{S_u} d|\nu| + \int_{S_u} \|\mu^i - \mu^e\|_1 + \int_{S_u} \|\mu^i - \sigma^i ds\|_1 + \|\mu^e - \sigma^e ds\|_1, $$

where the measure ν is defined by (5.10). It is easy to check that the functional H is homogeneous, so that, using [14], we can find a sequence ${\theta^i}^n, {\theta^e}^n$ in $L^1(S_u)$ such that

$$ {\theta^i}^n \xrightarrow[\mathbf{M}(S_u)]{} \mu^i, $$

$$ {\theta^e}^n \xrightarrow[\mathbf{M}(S_u)]{} \mu^e, $$

$$ \lim_n H(x, {\theta^i}^n, {\theta^e}^n, \sigma^i, \sigma^e) = H(x, \mu^i, \mu^e, \sigma^i, \sigma^e). $$

Consequently, the constructed sequence $(\sigma^n, {\theta^i}^n, {\theta^e}^n)$ permits us to get (5.27), which concludes the proof. ∎

5.4. The relaxation results.

PROPOSITION 5.4. *Let E_1 be the functional defined by* (5.11) *with hypotheses* (4.23)–(4.24), (4.2)–(4.3), (4.4)–(4.6), (4.7)–(4.8). *Then the relaxed functional of E_1 for the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak is*

(5.28)

$$
\begin{aligned}
&R(E_1) : \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u) \to R,\\
&R(E_1)(\sigma, \mu^i, \mu^e) = \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \sum_{j=1}^{2} \int_{\Omega^i \cup \Omega^e} \phi(D\sigma_j) + \int_\Omega c(x)\|\sigma\|^2\\
&\qquad + \int_{S_u} d|\nu| + \int_{S_u} \|\mu^i - \mu^e\|_1 + \int_{S_u} \|\mu^i - \sigma^i ds\|_1 + \|\mu^e - \sigma^e ds\|_1,\\
&\qquad \textit{where} \quad \nu = \frac{\mu^i + \mu^e}{2} \cdot n(u^+ - u^-) + h_2 ds.
\end{aligned}
$$

We can verify that we have $R(E_1) \leq E_1$ and that they are equal as soon as $\mu^e = \sigma^e ds$ and $\mu^i = \sigma^i ds$.

Proof. To simplify notations, we will note τ the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-strong $\times$ $\mathbf{M}(S_u)$-weak $\times$ $\mathbf{M}(S_u)$-weak and τ^d the topology $\mathbf{L^2}(\Omega^i \cup \Omega^e)$-weak $\times$ $\mathcal{C}^0(S_u)$-strong $\times$ $\mathcal{C}^0(S_u)$-strong. Notice that we will also use the notation M to denote a universal constant appearing in uniform bounds. This value may change from one line to another, but we will always write M.

We first remark that using Lemma 5.2 permits us to work on a more regular space, that is, with E_2. $R(E_2)$ is the relaxed functional of E_2 (or equivalently of E_1, thanks to Lemma 5.2) if and only if, for all $(\sigma, \mu^i, \mu^e) \in \mathbf{W^{1,1}}(\Omega^i \cup \Omega^e) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)$, we have the following two conditions:

(i) for all $(\sigma^n, {\mu^i}^n, {\mu^e}^n) \underset{\tau}{\to} (\sigma, \mu^i, \mu^e)$, then

$$\liminf_{n\to\infty} E_2(\sigma^n, {\mu^i}^n, {\mu^e}^n) \geq R(E_2)(\sigma, \mu^i, \mu^e); \tag{5.29}$$

(ii) there exists $(\sigma^n, {\mu^i}^n, {\mu^e}^n) \underset{\tau}{\to} (\sigma, \mu^i, \mu^e)$ such that

$$\limsup_{n\to\infty} E_2(\sigma^n, {\mu^i}^n, {\mu^e}^n) \leq R(E_2)(\sigma, \mu^i, \mu^e). \tag{5.30}$$

The purpose of the two steps below is to establish that $R(E_2) = \overline{E}_2$.

Step 1. This part is devoted to proving that

$$\liminf_{n\to\infty} E_2(\sigma^n, {\mu^i}^n, {\mu^e}^n) \geq R(E_2)(\sigma, \mu^i, \mu^e)$$

for all the sequences $(\sigma^n, {\mu^i}^n, {\mu^e}^n)$ converging to (σ, μ^i, μ^e).

Let $(\sigma, \mu^i, \mu^e) \in \mathbf{W^{1,1}}(\Omega^i \cup \Omega^e) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)$ and a sequence $(\sigma^n, {\mu^i}^n, {\mu^e}^n) \in \mathbf{BV}(\Omega) \times \mathbf{M}(S_u) \times \mathbf{M}(S_u)$ converging to (σ, μ^i, μ^e) for the τ-topology such that

$$E_2(\sigma^n, {\mu^i}^n, {\mu^e}^n) \leq M,$$

where M is a constant. Then, thanks to (4.3), we deduce that

$$\|\sigma^n\|_{BV(\Omega)} \leq M, \tag{5.31}$$

$$\text{on } S_u \text{ we have } \left\{ \begin{array}{l} \mu^{i^n} = \sigma^{i^n} ds, \\ \mu^{e^n} = \sigma^{e^n} ds, \end{array} \right. \quad \text{so} \quad \left\{ \begin{array}{l} \sigma^{i^n} ds \xrightarrow[\mathbf{M}(S_u)]{} \mu^i, \\ \sigma^{e^n} ds \xrightarrow[\mathbf{M}(S_u)]{} \mu^e. \end{array} \right. \tag{5.32}$$

Since the sequence (σ^n) is bounded in $BV(\Omega)$, we can deduce that there exists a measure $\tilde{\mu}$ such that

$$\sigma^n ds \xrightarrow[\mathcal{M}(\partial\Omega)]{} \tilde{\mu}. \tag{5.33}$$

Decomposing Ω permits us to write

$$\begin{aligned} E_2(\sigma^n, \mu^{i^n}, \mu^{e^n}) = & \int_\Omega |\sigma^n \cdot \nabla u + h_1| dx + \int_\Omega c(x) \|\sigma^n\|^2 dx \\ & + \sum_{j=1}^{2} \left(\int_{\Omega^i \cup \Omega^e} \phi(\|\nabla \sigma_j^n\|) dx \right) + \int_{S_u} \|\sigma^{i^n} - \sigma^{e^n}\|_1 ds \\ & + \int_{S_u} d|\nu^n|, \end{aligned}$$

where ν^n is given by (5.10). Thanks to the convergence properties, it is easy to check that

$$\begin{aligned} \liminf_{n\to\infty} \int_\Omega |\sigma^n \cdot \nabla u + h_1| dx + \int_\Omega c(x) \|\sigma^n\|^2 dx &\geq \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \int_\Omega c(x) \|\sigma\|^2 dx, \\ \liminf_{n\to\infty} \int_{S_u} \|\sigma^{i^n} - \sigma^{e^n}\|_1 ds &\geq \int_{S_u} \|\mu^i - \mu^e\|_1 ds, \\ \liminf_{n\to\infty} \int_{S_u} d|\nu^n| &\geq \int_{S_u} d|\nu|. \end{aligned} \tag{5.34}$$

By classical arguments and for a fixed j, we also have

$$\liminf_{n\to\infty} \int_{\Omega^i \cup \Omega^e} \phi(\|\nabla \sigma_j^n\|) dx \geq \int_{\Omega^i \cup \Omega^e} \phi(D\sigma_j) dx.$$

However, this minoration will not allow us to conclude anything because we need to be more precise. Using the notation proposed in Figure 5.1, and especially the decomposition $\Omega^i \cup \Omega^e = \Omega^{i,\alpha} \cup \Omega^{e,\alpha} \cup \Omega_{S_u}^{i,\alpha} \cup \Omega_{S_u}^{e,\alpha}$, we can write

$$\int_{\Omega^i \cup \Omega^e} \phi(\|\nabla {\sigma_j}^n\|) dx = \underbrace{\int_{\Omega^{i,\alpha} \cup \Omega^{e,\alpha}} \phi(\|\nabla {\sigma_j}^n\|) dx}_{\mathtt{A}^\alpha} + \underbrace{\int_{\Omega_{S_u}^{i,\alpha} \cup \Omega_{S_u}^{e,\alpha}} \phi(\|\nabla {\sigma_j}^n\|) dx}_{\mathtt{B}^\alpha}. \tag{5.35}$$

We study both parts separately.

Integral A^α. By classical arguments [25], we can write

$$\liminf_{n\to\infty} \int_{\Omega^{i,\alpha} \cup \Omega^{e,\alpha}} \phi(\|\nabla \sigma_j^n\|) dx \geq \int_{\Omega^{i,\alpha} \cup \Omega^{e,\alpha}} \phi(D\sigma_j) dx. \tag{5.36}$$

Integral B^α. Thanks to (4.6), we have

$$\int_{\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}}\phi(\|\nabla\sigma_j{}^n\|)dx \geq \int_{\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}}\phi^\infty(\|\nabla\sigma_j{}^n\|)dx - k|\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}|.$$

However, since ϕ^∞ is convex, we have

$$\phi^\infty(\|\nabla\sigma_j{}^n\|) = \sup_{q^{Gj}\in R^2} q^{Gj}\cdot\nabla\sigma_j^n - (\phi^\infty)^*(q^{Gj}) \ ,$$

where

$$(\phi^\infty)^*(x^*) = \begin{cases} 0 & \text{if } x^*\in dom(\phi^*) = \{x^*\in R^2/|x^*|<1\},\\ +\infty & \text{otherwise.}\end{cases}$$

So, for any function $q^{Gj}(x)$ such that $q^{Gj}(x)\in dom(\phi^*)$ a.e. $x\in\Omega$ with q^{Gj} in K defined by

$$\begin{aligned} K = \{q\in&\mathbf{L^2}(\Omega) \text{ such that}\\ &\|q^i\|_{L^\infty(\Omega)}\leq 1, \|q^e\|_{L^\infty(\Omega)}\leq 1,\\ &\operatorname{div}(q^i) \text{ and } \operatorname{div}(q^e)\in L^2(\Omega),\\ &q^i\cdot n|_{S_u} \text{ and } q^e\cdot n|_{S_u}\in\mathcal{C}^0(S_u),\\ &q\cdot n|_{\partial\Omega}=0\},\end{aligned} \tag{5.37}$$

we have

$$\int_{\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}}\phi^\infty(\|\nabla\sigma_j{}^n\|) \geq \int_{\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}} q^{Gj}\cdot\nabla\sigma_j^n. \tag{5.38}$$

The set K has been introduced so that all the integrals that we are going to write below are well defined. The only remaining problem is to estimate the limit of the right-hand side. To this end, we first integrate by parts the same term, but on Ω. We have

$$\begin{aligned}\int_\Omega q^{Gj}\cdot\nabla\sigma_j^n dx &= \int_{\Omega^i} q^{Gj}\cdot\nabla\sigma_j^n dx + \int_{\Omega^e} q^{Gj}\cdot\nabla\sigma_j^n dx\\ &= -\int_{\Omega^i\cup\Omega^e}\operatorname{div}(q^{Gj})\sigma_j^n dx + \int_{S_u}(q^{Gj}{}^i\sigma_j^{i\,n} - q^{Gj}{}^e\sigma_j^{e\,n})\cdot n ds + \int_{\partial\Omega} q^{Gj}\sigma_j^n\cdot n ds.\end{aligned}$$

Thanks to the strong convergence in $\mathbf{L^2}(\Omega^i\cup\Omega^e)$ of the sequence σ^n and to (5.32) (5.33), we have

$$\begin{aligned}&\lim_{n\to\infty}\int_\Omega q^{Gj}\cdot\nabla\sigma_j^n\\ &\quad= -\int_{\Omega^i\cup\Omega^e}\operatorname{div}(q^{Gj})\sigma_j dx + \int_{S_u}(q^{Gj}{}^i\mu^i{}_j - q^{Gj}{}^e\mu^e{}_j)\cdot n + \int_{\partial\Omega} q^{Gj}\tilde{\mu}_j\cdot n.\end{aligned} \tag{5.39}$$

Moreover, if we consider $q^{Gj}\cdot\nabla\sigma_j^n$ as a measure, for $\varphi\in\mathcal{C}^1_c(\Omega)$, we have by (4.11)

$$\langle q^{Gj}\cdot\nabla\sigma_j^n,\varphi\rangle = -\int_\Omega\operatorname{div}(q^{Gj})\sigma_j^n\varphi dx - \int_\Omega q^{Gj}\cdot\nabla\varphi\sigma_j^n dx.$$

When n tends to infinity, we have

$$\lim_{n\to\infty}\langle q^{Gj}\cdot\nabla\sigma_j^n,\varphi\rangle = -\int_\Omega \mathrm{div}(q^{Gj})\sigma_j\varphi - \int_\Omega q^{Gj}\cdot\nabla\varphi\sigma_j$$
$$= \langle q^{Gj}\cdot\nabla\sigma_j,\varphi\rangle.$$

The last result is the same as saying that the measure $q^{Gj}\cdot\nabla\sigma_j^n$ converges to $q^{Gj}\cdot\nabla\sigma_j$ for the topology $\mathcal{M}(\Omega)$weak$\star$. Since Ω is bounded, we can prove in fact that

$$\lim_{n\to\infty}\int_{\Omega^{i,\alpha}\cup\Omega^{e,\alpha}} q^{Gj}\cdot\nabla\sigma_j^n dx = \int_{\Omega^{i,\alpha}\cup\Omega^{e,\alpha}} q^{Gj}\cdot\nabla\sigma_j dx. \tag{5.40}$$

Consequently, substracting (5.40) from (5.39) permits us to write

$$\lim_{n\to\infty}\int_{\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}} q^{Gj}\cdot\nabla\sigma_j^n dx = -\int_{\Omega^i\cup\Omega^e}\mathrm{div}(q^{Gj})\sigma_j dx - \int_{\Omega^{i,\alpha}\cup\Omega^{e,\alpha}} q^{Gj}\cdot\nabla\sigma_j dx$$
$$+\int_{S_u}(q^{Gj^i}\mu^i{}_j - q^{Gj^e}\mu^e{}_j)\cdot n + \int_{\partial\Omega} q^{Gj}\tilde{\mu}_j\cdot n,$$

and, after integrating by part the term $\displaystyle\int_{\Omega^{i,\alpha}\cup\Omega^{e,\alpha}} q^{Gj}\cdot\nabla\sigma_j dx$, we can rewrite (5.38):

$$\lim_{n\to\infty}\int_{\Omega^{i,\alpha}_{S_u}\cup\Omega^{e,\alpha}_{S_u}}\phi(\|\nabla\sigma_j^n\|)dx \geq -\int_{\Omega^i\cup\Omega^e}\mathrm{div}(q^{Gj})\sigma_j dx + \int_{\Omega^{i,\alpha}\cup\Omega^{e,\alpha}}\mathrm{div}(q^{Gj})\sigma_j dx$$
$$+\int_{S_u}(q^{Gj^i}\mu^i{}_j - q^{Gj^e}\mu^e{}_j)\cdot n$$
$$-\int_{S_u^{i,\alpha}} q^{Gj^i}\sigma_j^i\cdot n ds - \int_{S_u^{e,\alpha}} q^{Gj^e}\sigma_j^e\cdot n ds$$
$$+\int_{\partial\Omega} q^{Gj}\cdot n(\tilde{\mu}_j - \sigma_j ds). \tag{5.41}$$

Now that we have found a minoration for integrals A^α (5.36) and B^α (5.41), we merge both results and let α tend to 0. The obtained result is

$$\begin{aligned}
&\liminf_{n\to\infty} E_2(\sigma^n,\mu^{i^n},\mu^{e^n})\\
&\quad\geq \int_\Omega|\sigma\cdot\nabla u + h_1|dx + \sum_{j=1}^2\int_{\Omega^i\cup\Omega^e}\phi(D\sigma_j) + \int_\Omega c(x)\|\sigma\|^2 + \int_{S_u} d|\nu| + \int_{S_u}\|\mu^i-\mu^e\|_1\\
(\blacklozenge)\quad &+\sum_{j=1}^2\int_{S_u} q^{Gj^i}\cdot n(\mu^i{}_j - \sigma_j^i ds) - \sum_{j=1}^2\int_{S_u} q^{Gj^e}\cdot n(\mu^e{}_j - \sigma_j^e ds)\\
(\Diamond)\quad &+\sum_{j=1}^2\int_{\partial\Omega} q^{Gj}\cdot n(\tilde{\mu}_j - \sigma_j ds),
\end{aligned} \tag{5.42}$$

where $\nu = \dfrac{\mu^i+\mu^e}{2}\cdot n(u^+-u^-) + h_2 ds$. Since this inequality is true for all $\mathbf{q}^G = (q^{G1},q^{G2})\in K\times K$, it is still true when we take the supremum in $\mathbf{q}^G$. This supremum is taken for $q^{G1},q^{G2}\in K$ defined by (5.37). Next, we introduce the set

$$C(x) = \mathrm{closure}\{(q^i(x)\cdot n(x), q^e(x)\cdot n(x)),\quad (q^i,q^e)\in K\},$$

which can be rewritten as

$$
\begin{aligned}
C(x) = \text{closure}\{&(z^i,z^e) \in R^2 \text{ such that} \\
&\exists(\varphi^i,\varphi^e) \in [\mathcal{C}^0(S_u)]^2/\varphi^i(x) = z^i, \varphi^e(x) = z^e, \\
&\exists(q^i,q^e), \|q^i\|_\infty \le 1, \|q^e\|_\infty \le 1, \text{div}(q^i) \text{ and } \text{ div}(q^e) \in L^2(\Omega), \\
&q^i(x)\cdot n(x) = \varphi^i(x), q^e(x)\cdot n(x) = \varphi^e(x)\ ,\ x \in S_u \\
&q(x)\cdot n = 0 \quad x \in \partial\Omega\}.
\end{aligned}
$$

To compute the supremum of (5.42), we consider in (5.42) only the term noted with the symbol (⧫). Notice that the term noted with the symbol (◊) will not appear in the minimization thanks to the definition of the set K, where we have imposed $q\cdot n|_{\partial\Omega} = 0$. If we note

$$
W = \begin{pmatrix} (\mu^i{}_1 - \sigma^i_1, \mu^e{}_1 - \sigma^e_1) \\ (\mu^i{}_2 - \sigma^i_2, \mu^e{}_2 - \sigma^e_2) \end{pmatrix},
$$

we claim that

$$
\begin{aligned}
&\sup_{(q^{G1},q^{G2})\in K^2} \left\{ \sum_{j=1}^{2} \int_{S_u} q^{Gj\,i}\cdot n(\mu^i{}_j - \sigma^i_j ds) - \sum_{j=1}^{2}\int_{S_u} q^{Gj\,e}\cdot n(\mu^e{}_j - \sigma^e_j ds) \right\} \\
&= \sup_{Z} \int_{S_u} \sum_{j=1}^{2} Z^t_j \cdot W_j, \quad \text{where } Z = \begin{pmatrix} (z^i_1, z^e_1) \\ (z^i_2, z^e_2) \end{pmatrix} \in C(x)^2 \qquad (5.43) \\
&= \int_{S_u} \|\mu^i - \sigma^i ds\|_1 + \|\mu^e - \sigma^e ds\|_1. \qquad (5.44)
\end{aligned}
$$

Equality (5.43) corresponds to the permutation of the supremum. It is based on techniques developed in [15, 17]. We then need to express that supremum giving the expression (5.44) [14, 13]. We refer to [35] for the complete proof.

Finally, taking the supremum in (5.42) with respect to $q^{Gj\,i}$,$q^{Gj\,e}$ and using (5.44) permits us to have

$$
\liminf_{n\to\infty} E_2(\sigma^n, \mu^{i^n}, \mu^{e^n}) \ge \overline{E}_2(\sigma, \mu^i, \mu^e), \qquad (5.45)
$$

where $\overline{E}_2$ has been previously defined in (5.17). The functional $\overline{E}_2$ is then a candidate to be the relaxed functional of E_2. It remains to show the second condition (5.30) (with $R(E_2) = \overline{E}_2$), which is the aim of Step 2.

Step 2. The way to demonstrate (5.30) is based on the following assertion. Showing (5.30) is equivalent to proving that

$$
\begin{aligned}
&\text{for all } \ (f,\varphi^i,\varphi^e) \in \mathbf{L}^\infty(\Omega^i \cup \Omega^e) \times \mathcal{C}^0(S_u) \times \mathcal{C}^0(S_u) \\
&\text{there exists } (f^n, \varphi^{i^n}, \varphi^{e^n}) \underset{\tau^d}{\to} (f,\varphi^i,\varphi^e) \ \text{ such that} \\
&\liminf_{n\to\infty} E_2^*(f^n, \varphi^{i^n}, \varphi^{e^n}) \ge R(E_2)^*(f,\varphi^i,\varphi^e), \qquad (5.46)
\end{aligned}
$$

where the superscript $*$ denotes the conjugate functionals. This result is due to [8] (see also [13], where this idea has been used).

Naturally, the difficulty is to compute the conjugate functional of E_2 and $\overline{E_2}$. This is done in Lemmas A.1 and A.2 of the appendix. We have shown that

$$E_2^*(f,\varphi^i,\varphi^e) = \inf_{q\in\mathcal{A}(f,\varphi^i,\varphi^e)} J(q) \ ,$$

where the minimum is computed for $q=(q^\sigma,q^{G1},q^{G2},q^u,q^{T1},q^{T2},q^{S_u})$ in $(\mathbf{L}^\infty(\Omega^\mathbf{i}\cup\Omega^\mathbf{e}))^3 \times L^\infty(\Omega^i\cup\Omega^e)\times(\mathbf{L}^\infty(\mathbf{S_u}))^2\times L^\infty(S_u)$ verifying the conditions (A.2)–(A.10) (which defines the set $\mathcal{A}(f,\varphi^i,\varphi^e)$) and

$$\overline{E_2}^*(f,\varphi^i,\varphi^e) = \inf_{\overline{q}\in\overline{\mathcal{A}}(f,\varphi^i,\varphi^e)} \overline{J}(\overline{q})$$

when the minimum is computed for $\overline{q}=(\overline{q}^\sigma,\overline{q}^{G1},\overline{q}^{G2},\overline{q}^u,\overline{q}^{T1},\overline{q}^{T2},\overline{q}^{\mu^i},\overline{q}^{\mu^e},\overline{q}^{S_u})$ in $(\mathbf{L}^\infty(\Omega^\mathbf{i}\cup\Omega^\mathbf{e}))^3\times L^\infty(\Omega^i\cup\Omega^e)\times(\mathbf{L}^\infty(\mathbf{S_u}))^4\times L^\infty(S_u)$ verifying conditions (A.13)–(A.24) (which defines the set $\overline{\mathcal{A}}(f,\varphi^i,\varphi^e)$). For more details about the definitions of $J,\overline{J},\mathcal{A}(f,\varphi^i,\varphi^e)$ and $\overline{\mathcal{A}}(f,\varphi^i,\varphi^e)$, we refer to Lemmas A.1 and A.2.

Let $(f^n,{\varphi^i}^n,{\varphi^e}^n)$ be a sequence such that

$$\liminf_{n\to\infty} E_2^*(f^n,{\varphi^i}^n,{\varphi^e}^n) \le M, \tag{5.47}$$

where M is a constant. Then, for each n and using the definition of the conjugate function associated with E_2 (Lemma A.1), there exists a $q^n\in\mathcal{A}(f^n,{\varphi^i}^n,{\varphi^e}^n)$ so that

$$E_2^*(f^n,{\varphi^i}^n,{\varphi^e}^n) \ge J(q^n) - \frac{1}{n}. \tag{5.48}$$

Since $E_2^*(f^n,{\varphi^i}^n,{\varphi^e}^n)$ is uniformly bounded thanks to (5.47), it is easy to check that we can find an element $q\in\mathcal{A}(f,\varphi^i,\varphi^e)$ such that the sequence q^n converges to q for the weak topology of this space, that is to say, the topology

$$(\mathbf{L}^\infty(\Omega^\mathbf{i}\cup\Omega^\mathbf{e}))^3\text{weak}\times L^\infty(\Omega^i\cup\Omega^e)\text{weak}\times(\mathbf{L}^\infty(\mathbf{S_u}))^2\text{weak}\times L^\infty(S_u)\text{weak}.$$

As the function J is lower semicontinuous for this topology, we have

$$\liminf_{n\to\infty} J(q^n) \ge J(q) \ge \inf_{q\in\mathcal{A}(f,\varphi^i,\varphi^e)} J(q) \ . \tag{5.49}$$

Now, let us define the application T by

$$\begin{aligned} &T:\mathcal{A}(f,\varphi^i,\varphi^e)\to\overline{\mathcal{A}}(f,\varphi^i,\varphi^e),\\ &T(q)=\overline{q}, \end{aligned} \tag{5.50}$$

where $\overline{q}$ is defined by:

$$\begin{aligned}
\overline{q}^{\sigma} &= q^{\sigma},\\
\overline{q}^{Gj} &= q^{Gj}, \quad j = 1, 2,\\
\overline{q}^{u} &= q^{u},\\
\overline{q}^{S_u} &= q^{S_u},\\
\overline{q}^{Ti} &= \frac{1}{2} q^{S_u} \cdot n(u^+ - u^-) + q^{Ti},\\
\overline{q}^{Te} &= \frac{1}{2} q^{S_u} \cdot n(u^+ - u^-) + q^{Te},\\
\overline{q}^{\mu^i} &= \varphi^i - \frac{1}{2} q^{S_u} \cdot n(u^+ - u^-),\\
\overline{q}^{\mu^e} &= \varphi^e - \frac{1}{2} q^{S_u} \cdot n(u^+ - u^-).
\end{aligned}$$

An easy computation permits us to see that if q belongs to $\mathcal{A}(f, \varphi^i, \varphi^e)$, then $T(q)$ belongs to $\overline{\mathcal{A}}(f, \varphi^i, \varphi^e)$. Moreover, we can observe that

$$J(q) = \overline{J}(T(q)).$$

Consequently, using (5.48)–(5.49), the definition of the function T and Lemma A.2, we have

$$\begin{aligned}
\liminf_{n\to\infty} E_2^*(f^n, {\varphi^i}^n, {\varphi^e}^n) &\geq \inf_{q \in \mathcal{A}(f,\varphi^i,\varphi^e)} J(q) = \inf_{\overline{q} \in \overline{\mathcal{A}}(f,\varphi^i,\varphi^e) \cap Im(T)} \overline{J}(\overline{q})\\
&\geq \inf_{\overline{q} \in \overline{\mathcal{A}}(f,\varphi^i,\varphi^e)} \overline{J}(\overline{q}) = \overline{E}_2^*(f, \varphi^i, \varphi^e),
\end{aligned} \tag{5.51}$$

which is exactly the statement (5.46).

Conclusion. As we can observe, the functional $\overline{E}_2$ complies with conditions (5.45) and (5.51), that is to say, (5.29)–(5.30) (or, equivalently, (5.29)–(5.46)). As a conclusion, we have

$$R(E_2) = \overline{E}_2,$$

which is the desired statement. □

PROPOSITION 5.5. *The relaxed functional noted $R(E)$ of the functional E defined by* (5.3) *is given by*

(5.52)

$$R(E) : \mathbf{BV}(\Omega) \to R,$$

$$R(E)(\sigma) = \int_\Omega |\sigma \cdot \nabla u + h_1| dx + \sum_{j=1}^{2} \int_{\Omega^i \cup \Omega^e} \phi(D\sigma_j) + \int_\Omega c(x) \|\sigma\|^2 + \int_{S_u} \beta(x, \sigma^i, \sigma^e),$$

where

$$\beta(x, \lambda, \theta) = \operatorname{Inf} \left\{ |\lambda - s| + |\theta - t| + |s - t| + \left| \frac{s+t}{2} \cdot n(x)(u^+ - u^-) + h_2(x) \right| \ : \right. \tag{5.53}$$

$$\left. (s, t) \in R^N \times R^N \right\}.$$

Proof. This proposition is a direct consequence of Proposition 5.4, and we will just sketch the proof. Let us define

$$G(\sigma) = \inf_{(\mu^i,\mu^e)\in\mathcal{M}(\Omega)} \overline{E_1}(\sigma,\mu^i,\mu^e) \ .$$

By classical arguments, we prove that the functional G is lower semicontinuous, less than E, and also greater that $R(E)$; so in fact

$$G(\sigma) = R(E)(\sigma).$$

We deduce the final result from a Rockafellar theorem [42, 44] which permits to permute the infimum and the integral. □

6. Existence for the relaxed functional.

PROPOSITION 6.1. *Let $R(E)$ be defined by* (5.52) *and* (**??**), *where u verifies hypotheses* (4.23)–(4.24), *$\phi(\cdot)$ satisfies* (4.2)–(4.3), (4.4)–(4.6), *and $c(x)$ is a function verifying* (4.7)–(4.8). *Then the problem* $\mathrm{Inf}\{R(E)(\sigma) : \sigma \in \mathbf{BV}(\Omega)\}$ *admits a solution in* $\mathbf{BV}(\Omega)$.

Proof. The functional $R(E)$ is a convex function of measures which is lower semicontinuous by construction. Moreover, it is coercive, so we can uniformly bound minimizing sequences and deduce by classical arguments the existence of a solution. □

The above theorem proves an existence result for the relaxed functional associated to the optical flow problem. The main difficulty came from the product $(\sigma \cdot Du)$, for which we found an explicit integral representation. It will be interesting to study more general functionals involving terms of the form $f((\sigma \cdot Du))$. This question will be considered in a forthcoming paper.

Another challenging problem is the numerical analysis of these abstract results. This induces several difficulties. One of the first is to characterize the solution. No Euler equations can be written, but some partial answers have been given, using variational [4] or dual [51] formulations. Then it will be necessary to propose some suitable discretizations to take into account the discontinuities of the solution. These problems will be considered in the future.

Appendix A. The dual functionals E_2^* and $\overline{E}_2^*$. We give in the two lemmas below the detailed expressions of the dual functions associated to E_2 and $\overline{E}_2$.

LEMMA A.1. *Let E_2 be given by* (5.16). *Its dual functional is defined by*

$$\text{(A.1)} \qquad \begin{aligned} &E_2^* : \mathbf{L}^\infty(\Omega^i \cup \Omega^e) \times \mathcal{C}^0(S_u) \times \mathcal{C}^0(S_u) \to R, \\ &E_2^*(f,\varphi^i,\varphi^e) = \inf_{q\in\mathcal{A}(f,\varphi^i,\varphi^e)} J(q), \end{aligned}$$

where the infimum is taken over $q = (q^\sigma, q^{G1}, q^{G2}, q^u, q^{Ti}, q^{Te}, q^{S_u})$ belonging to $(\mathbf{L}^\infty(\Omega^\mathbf{i} \cup \Omega^\mathbf{e}))^3 \times L^\infty(\Omega^i \cup \Omega^e) \times (\mathbf{L}^\infty(\mathbf{S_u}))^2 \times L^\infty(S_u)$ and complying the following

conditions:

$$|q^u| \le 1 \quad \textit{a.e. on } \Omega, \tag{A.2}$$

$$|q^{Gj}| \le 1, j = 1, 2, \quad \textit{a.e. on } \Omega, \tag{A.3}$$

$$|q^{S_u}| \le 1 \quad \textit{a.e. on } S_u, \tag{A.4}$$

$$|q^{Ti} + \varphi^i| \le 1 \quad \textit{a.e. on } S_u, \tag{A.5}$$

$$q^u \nabla u + q^\sigma - \operatorname{div}(\mathbf{q}^{\mathbf{G}}) - f = 0 \quad \textit{on } \Omega, \tag{A.6}$$

$$\frac{1}{2} q^{S_u} n(u^+ - u^-) + q^{Ti} + \mathbf{q}^{\mathbf{G}^i} n = 0 \quad \textit{on } S_u, \tag{A.7}$$

$$\frac{1}{2} q^{S_u} n(u^+ - u^-) + q^{Te} + \mathbf{q}^{\mathbf{G}^e} n = 0 \quad \textit{on } S_u, \tag{A.8}$$

$$q^{Ti} + \varphi^i + q^{Te} + \varphi^e = 0 \quad \textit{on } S_u, \tag{A.9}$$

$$\mathbf{q}^{\mathbf{G}} n = 0 \quad \textit{on } \partial\Omega, \tag{A.10}$$

where $\mathbf{q}^{\mathbf{G}}$ *is the matrix defined by* $\mathbf{q}^{\mathbf{G}} = \begin{pmatrix} q^{G1^T} \\ q^{G2^T} \end{pmatrix}$ *and where the function* J *is defined by*

$$J(q) = \int_\Omega q^u h_1 dx + \int_\Omega \frac{1}{4c(x)^2} |q^\sigma|^2 dx + \sum_{j=1}^{2} \int_\Omega \phi^*(|q^{Gj}|) dx + \int_{S_u} q^{S_u} h_2 ds. \tag{A.11}$$

LEMMA A.2. *Let* $\overline{E}_2$ *given by* (5.17). *Its dual functional is defined by*

$$\begin{aligned} &\overline{E}_2^* : \mathbf{L}^\infty(\Omega^i \cup \Omega^e) \times \mathcal{C}^0(S_u) \times \mathcal{C}^0(S_u) \to R, \\ &\overline{E}_2^*(f, \varphi^i, \varphi^e) = \inf_{\overline{q} \in \overline{\mathcal{A}}(f, \varphi^i, \varphi^e)} \overline{J}(\overline{q}) \ , \end{aligned} \tag{A.12}$$

where the infimum is taken over $\overline{q} = (\overline{q}^\sigma, \overline{q}^{G1}, \overline{q}^{G2}, \overline{q}^u, \overline{q}^{Ti}, \overline{q}^{Te}, \overline{q}^{\mu^i}, \overline{q}^{\mu^e}, \overline{q}^{S_u})$ *belonging to* $(\mathbf{L}^\infty(\Omega^{\mathbf{i}} \cup \Omega^{\mathbf{e}}))^3 \times L^\infty(\Omega^i \cup \Omega^e) \times (\mathbf{L}^\infty(\mathbf{S_u}))^4 \times L^\infty(S_u)$ *and complying to the following conditions:*

$$|\overline{q}^u| \le 1 \quad \textit{a.e. on } \Omega, \tag{A.13}$$

$$|\overline{q}^{Gj}| \le 1 \quad (j = 1, 2) \quad \textit{a.e. on } \Omega, \tag{A.14}$$

$$|\overline{q}^{S_u}| \le 1 \quad \textit{a.e. on } S_u, \tag{A.15}$$

$$|\overline{q}^{Ti} + \overline{q}^{\mu^i}| \le 1 \quad \textit{a.e. on } S_u, \tag{A.16}$$

$$q^{Ti} \quad \textit{and} \quad q^{Te} \in \mathcal{C}(x) \quad \textit{a.e. on } S_u, \tag{A.17}$$

$$\overline{q}^u \cdot \nabla u + \overline{q}^\sigma - div(\overline{\mathbf{q}}^{\mathbf{G}}) - f = 0 \quad \textit{on } \Omega, \tag{A.18}$$

$$q^{Ti} + \overline{\mathbf{q}}^{\mathbf{G}^i} : n = 0 \quad \textit{on } S_u, \tag{A.19}$$

$$\overline{q}^{Te} + \overline{\mathbf{q}}^{\mathbf{G}^e} : n = 0 \quad \textit{on } S_u, \tag{A.20}$$

$$\frac{1}{2} \overline{q}^{S_u} n(u^+ - u^-) + \overline{q}^{\mu^i} - \varphi^i = 0 \quad \textit{on } S_u, \tag{A.21}$$

$$\frac{1}{2} \overline{q}^{S_u} n(u^+ - u^-) + \overline{q}^{\mu^e} - \varphi^e = 0 \quad \textit{on } S_u, \tag{A.22}$$

$$\overline{q}^{Ti} + \overline{q}^{\mu^i} + \overline{q}^{Te} + \overline{q}^{\mu^e} = 0 \quad \textit{on } S_u, \tag{A.23}$$

$$\overline{\mathbf{q}}^{\mathbf{G}} : n = 0 \quad \textit{on } \partial\Omega, \tag{A.24}$$

where $\overline{\mathbf{q}}^{\mathbf{G}}$ *is the matrix defined by* $\overline{\mathbf{q}}^{\mathbf{G}} = \begin{pmatrix} \overline{q}^{G1^T} \\ \overline{q}^{G2^T} \end{pmatrix}$ *and where the function* $\overline{J}$ *is defined by*

$$\overline{J}(\overline{q}) = \int_\Omega \overline{q}^u h_1 dx + \int_\Omega \frac{1}{4c(x)^2} |\overline{q}^\sigma|^2 dx + \sum_{j=1}^{2} \int_\Omega \phi^*(|\overline{q}^{Gj}|) dx + \int_{S_u} \overline{q}^{S_u} h_2 ds. \tag{A.25}$$

Proof of Lemmas A.1 *and* A.2. We refer to [35] for the complete proof, which is mainly technical. To get that result, we used classical techniques developed in [21], Rockafellar's theorem, and suitable choices of dual variables $q, \overline{q}$ to simplify calculus. □

Acknowledgments. We thank M. Bellieud, G. Bouchitté, and G. Buttazzo for their useful suggestions about relaxation results.

REFERENCES

[1] R. Acart and C.R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

[2] L. Ambrosio, *A compactness theorem for a new class of functions of bounded variation*, Boll. Uni. Mat. Ital. B(7), 3 (1989), pp. 857–881.

[3] G. Anzellotti, *Pairings between measures and bounded functions and compensated compactness*, Ann. Mat. Pura Appl. (4), 135 (1983), pp. 293–318.

[4] G. Anzellotti, *The Euler equation for functionals with linear growth*, Trans. Amer. Math. Soc., 290 (1985), pp. 483–501.

[5] G. Aubert, M. Barlaud, L. Blanc-Feraud, and P. Charbonnier, *Deterministic edge-preserving regularization in computed imaging*, IEEE Trans. Imag. Process., 5 (1997), pp. 298–311.

[6] G. Aubert, R. Deriche, and P. Kornprobst, *Computing optical flow via variational techniques*, SIAM J. Appl. Math., to appear.

[7] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[8] D. Aze, *Convergence of dual variables in problems of transmission through thin layers by epi-convergence methods*, Ricerche Mat., 35 (1986), pp. 125–159.

[9] J.L. Barron, D.J. Fleet, and S.S. Beauchemin, *Performance of optical flow techniques*, Internat. J. Computer Vision, 12 (1994), pp. 43–77.

[10] A.C. Barroso, G. Bouchitté, G. Buttazzo, and I. Fonseca, *Relaxation of bulk and interfacial energies*, Arch. Rational Mech. Anal., 135 (1996), pp. 107–173.

[11] G. Bouchitté, I. Fonseca, and L. Mascarenhas, *A global method for relaxation*, Technical report, Université de Toulon et du Var, May 1997.

[12] G. Bouchitté and G. Dal Maso, *Integral representation and relaxation of convex local functionals on $BV(\Omega)$*, Ann. Scuola Norm. Sup. Pisa Cl. Ser. (4), 20 (1993), pp. 483–533.

[13] G. Bouchitté and P. Suquet, *Homogenization, plasticity and yield design*, in Proceedings International Cent. Theor. Phys. Workshop, Composite Media and Homogenization Theory 5, G. Dal Maso and G. Dell'Antonio, eds., Birkhauser, Trieste, 1991, pp. 107–133.

[14] G. Bouchitté and M. Valadier, *Integral representation of convex functions on a space of measures*, J. Funct. Anal., 80 (1988), pp. 398–420.

[15] G. Bouchitté and M. Valadier, *Multi-fonctions s.c.i. et régularisée s.c.i. essentielle. fonctions de mesure dans le cas sous linéaire*, in Proceedings Congrès Franco-Québecois Analyse non linéaire, Gauthier Villars, Paris, 1989.

[16] G. Buttazzo, *Semicontinuity, relaxation and integral representation in the calculus of variations*, Longman Scientific and Technical, Harlow, UK, 1989.

[17] C. Castaing and M. Valadier, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Math. 500, Springer-Verlag, Berlin, 1977.

[18] I. Cohen, *Nonlinear variational method for optical flow computation*, in Proceedings Eighth SCIA, Vol. 1, Springer-Verlag, New York, 1993, pp. 523–530.

[19] F. Demengel and R. Temam, *Convex functions of a measure and applications*, Indiana Univ. Math. J., 33 (1984), pp. 673–709.

[20] R. Deriche and O. Faugeras, *Les EDP en traitement des images et vision par ordinateur*, Traitement du Signal, 13 (1996), pp. 551–577.
[21] I. Ekeland and R. Temam, *Analyse convexe et problèmes variationnels*, Etudes mathematiques, Dunod; Gauthier–Villars, Paris–Bruxelles–Montreal, 1974.
[22] L.C. Evans and R.F. Gariepy, *Measure Theory and Fine Properties of Functions*, CRC Press, Boca Raton, FL, 1992.
[23] H. Federer, *Geometric Measure Theory*, Classics Math., Vol. 153, Springer-Verlag, Berlin, 1969.
[24] D.J. Fleet and A.D. Jepson, *Computation of component image velocity from local phase information*, Internat. J. Comput. Vision, 5 (1990), pp. 77–104.
[25] E. Giusti, *Minimal Surfaces and Functions of Bounded Variation*, Birkhäuser, Cambridge, MA, 1984.
[26] C. Goffman and J. Serrin, *Sublinear functions of measures and variational integrals*, Duke Math J., 31 (1964), pp. 159–178.
[27] F. Heitz and P. Bouthemy, *Multimodal estimation of discontinuous optical flow using Markov random fields*, IEEE Trans. Pattern Analysis and Machine Intelligence, 15 (1993), pp. 1217–1232.
[28] B.K. Horn, *Robot Vision*, MIT Press, Cambridge, MA, 1986.
[29] B.K. Horn and B.G. Schunk, *Determining optical flow*, Artificial Intelligence, 17 (1981), pp. 185–203.
[30] J.J. Koenderink, *Optic Flow*, Vision Research, 26 (1986), pp. 161–180.
[31] J.J. Koenderink and A.J. van Doorn, *Invariant properties of the motion parallax field due to the movement of rigid bodies relative to an observer*, Optica Acta, 22 (1975), pp. 717–723.
[32] J.J. Koenderink and A.J. van Doorn, *How an ambulant observer can construct a model of the environment from the geometrical structure of the visual inflow*, in Kybernetik, G. Hauske and E. Butenandt, eds., Oldenburg, Muenchen, 1978.
[33] J.J. Koenderink and A.J. van Doorn, *Affine structure from motion*, J. Optical Soc. America, A8 (1991), pp. 377–385.
[34] J.J. Koenderink and A.J. van Doorn, *Dynamic shape*, Biological Cybernetics, 53 (1986), pp. 383–396.
[35] P. Kornprobst, *Contributions à la restauration d'images et à l'analyse de séquences: Approches variationnelles et equations aux dérivées partielles*, Ph.D. thesis, Université de Nice-Sophia, Antipolis, France, 1998.
[36] G. Dal Maso, *An introduction to* Γ*-convergence*, Progress in Nonlinear Differential Equations and their Applications, Birkhäuser, Cambridge, MA, 1993.
[37] H.H. Nagel, *Direct estimation of optical flow and of its derivatives*, in Artificial and Biological Vision Systems, G.A. Orban and H.-H. Nagel, eds., Basic Research Series, Springer-Verlag, Berlin, 1992, pp. 191–224.
[38] H.H. Nagel and W. Enkelmann, *An investigation of smoothness constraint for the estimation of displacement vector fields from images sequences*, IEEE Trans. Pattern Analysis and Machine Intelligence, 8 (1986), pp. 565–593.
[39] M. Orkisz and P. Clarysse, *Estimation du flot optique en présence de discontinuités: Une revue*, Traitement du Signal, 13 (1996), pp. 489–513.
[40] M. Otte and H.H. Nagel, *Optical flow estimation: Advances and comparisons*, Computer Vision–ECCV'94, Vol.I, in Proceedings of the Third European Conference, Lecture Notes in Comput. Sci. 800, Computer Vision, Springer-Verlag, Berlin, 1994, pp. 51–70.
[41] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck, *Determination of optical flow and its discontinuities using non-linear diffusion*, Computer Vision—ECCV'94, Vol. II, in Proceedings of the Third European Conference, Lecture Notes in Comput. Sci. 801, Springer-Verlag, Berlin, 1994, pp. 295–304.
[42] R.T. Rockafellar, *Integrals which are convex functionals*, Pacific J. Math., 24 (1968), pp. 525–539.
[43] R.T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.
[44] R.T. Rockafellar, *Integral which are convex functionals* II, Pacific J. Math., 39 (1971), pp. 439–469.
[45] L. Rudin, S. Osher, and E. Fatemi, *Nonlinear total variation based noise removal algorithms*, Phys. D, 60 (1992), pp. 259–268.
[46] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1966.
[47] R. Temam, *Dual variational principles in mechanics and physics*, in Semi-Infinite Programming and Applications, A.V. Fiacco and K.O. Kortanek, eds., Lecture Notes in Economics and Mathematical Systems 215, Springer-Verlag, Berlin, 1981.
[48] A.N. Tikhonov and V.Y. Arsenin, *Solutions of Ill-Posed Problems*, Winston and Sons, Wash-

ington, DC, 1977.
[49] A. Verri, F. Girosi, and V. Torre, *Differential techniques for optical flow*, J. Optical Soc. America, A7 (1990), pp. 912–922.
[50] A. Verri and T. Poggio, *Motion field and optical flow: Qualitative properties*, IEEE Trans. Pattern Analysis and Machine Intelligence, 11 (1989), pp. 490–498.
[51] L. Vese, *Problèmes variationnels et EDP pour l'analyse d'images et l'évolution de courbes*, Ph.D. thesis, Université de Nice-Sophia, Antipolis, France, November 1996.
[52] A.I. Vol'pert, *The spaces BV and quasilinear equations*, Math. USSR-Sbornik, 2 (1967), pp. 225–267.
[53] W.P. Ziemer, *Weakly Differentiable Functions*, Springer-Verlag, Berlin, 1989.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1309–1342

REPRESENTATION OF WEAK LIMITS AND DEFINITION OF NONCONSERVATIVE PRODUCTS*

PHILIPPE G. LEFLOCH† AND ATHANASIOS E. TZAVARAS‡

Abstract. The goal of this article is to show that the notion of generalized graphs is able to represent the limit points of the sequence $\{g(u_n)\,du_n\}$ in the weak-$\star$ topology of measures when $\{u_n\}$ is a sequence of continuous functions of uniformly bounded variation. The representation theorem induces a natural definition for the nonconservative product $g(u)\,du$ in a BV context. Several existing definitions of nonconservative products are then compared, and the theory is applied to provide a notion of solutions and an existence theory to the Riemann problem for quasi-linear, strictly hyperbolic systems.

Key words. nonconservative products, hyperbolic systems, shock wave, self-similar solution

AMS subject classifications. Primary, 35L60; Secondary, 28A75

PII. S0036141098341794

1. Introduction.

The objective of this article is to present a theoretical frame for the definition and properties of nonconservative products in one space dimension. The issue of defining nonconservative products appears with Volpert's chain rule [31] for BV functions in several space dimensions. It is a central problem for defining a notion of weak solutions for a general quasi-linear hyperbolic system

$$\partial_t u + A(u)\,\partial_x u = 0, \qquad u(x,t) \in \mathbb{R}^N, \quad x \in \mathbb{R},\ t > 0. \tag{1.1}$$

Such systems appear in several models of the engineering and physics literature, e.g., [5, 8, 23, 24, 25, 28]. The origin of the nonconservative terms is usually a simplifying modeling assumption or a closure hypothesis. If (1.1) is conservative, i.e., $A(u) = \nabla F(u)$ for some $F : \mathbb{R}^N \to \mathbb{R}^N$, then weak solutions are defined in the sense of distributions. In the general case, however, the term $A(u)\,\partial_x u$ will contain products of discontinuous functions with measures, and its definition is not obvious. At present, successful definitions exist in the one-space dimensional BV framework by LeFloch [14, 15], Dal Maso, LeFloch, and Murat [10] and Raymond [27]. The definition in [10] is based on a family of Lipschitz paths, is stable under weak convergence, and leads to a solution of the Riemann problem in the class of genuinely nonlinear, strictly hyperbolic systems with Riemann data that are sufficiently close. It has prompted investigations on existence of weak solutions to (1.1), LeFloch and Liu [17], and on convergence of numerical schemes, Hou and LeFloch [12]. The concept of extended graphs is used in [27] to provide a general definition that is stable under weak convergence.

*Received by the editors July 13, 1998; accepted for publication February 15, 1999; published electronically October 8, 1999. This work was supported in part by TMR project HCL ERBFMRXCT960033.

http://www.siam.org/journals/sima/30-6/34179.html

†Centre de Mathématiques Appliquées and Centre National de la Recherche Scientifique, UA 756, Ecole Polytechnique, 91128 Palaiseau, France (lefloch@cmap.polytechnique.fr). The research of this author was partially supported by the Centre National de la Recherche Scientifique and by the National Science Foundation via a Faculty Early Career Development award (CAREER).

‡Department of Mathematics, University of Wisconsin, Madison, WI 53706, and Institute of Applied and Computational Mathematics, FORTH, 711 10 Heraklion, Crete (tzavaras@math.wisc.edu). This research was completed while this author was visiting the Centre de Mathématiques Appliquées of Ecole Polytechnique and was partially supported by the National Science Foundation and the Office of Naval Research.

Related issues appear in studies of transport equations with discontinuous coefficients (e.g., LeFloch [15, 16], LeFloch and Xin [20], Poupaud and Rascle [22], Bouchut and James [4]) and in minimization of certain types of functionals in the space of functions of bounded variation (e.g., Aviles and Giga [1], Raymond and Seghir [26]). The reader is referred to Colombeau [6], Colombeau and Leroux [7] for a theory of nonconservative products in a weaker functional framework.

Let $g : \mathbb{R}^N \to \mathbb{R}^N$ be a continuous function and $u : [a, b] \to \mathbb{R}^N$ be a function of bounded variation. Our scope is to provide a justifiable definition for the inner product of $g(u)$ and $\frac{du}{dx}$, formally given by $g(u)\frac{du}{dx} = \sum_{i=1}^N g^i(u)\,\frac{du^i}{dx}$. This definition will be suggested by a representation theory of the limit points of sequences $\{g(u_n)\frac{du_n}{dx}\}$ in weak topologies when the functions u_n are *smooth*. This viewpoint reflects the premise that (1.1) arises in the limit of regularized problems as the dissipative mechanisms, such as viscosity or relaxation time, tend to zero. Accordingly, the nonconservative product will appear as a limit of regularized sequences.

If u is a *continuous* BV function, there is a natural definition of the product $\mu = g(u)\,\frac{du}{dx}$ as a Radon measure on $\mathcal{C}[a, b]$. This is done by setting

$$\langle \mu, \theta \rangle = \int_{[a,b]} \theta(x) g(u(x))\, du(x), \quad \theta \in \mathcal{C}[a, b], \tag{1.2}$$

where the right-hand side is viewed as a Borel–Stieltjes integral relative to the (vector-valued, signed) measure generated by $u \in \mathcal{C} \cap BV$. This definition is appropriate when u is continuous. If u has discontinuities, definition (1.2) is "not stable," because the integral $\int f du$, for $f \in L^1(du)$, changes values when changing f at the points of discontinuity of u.

Consider a sequence $\{u_n\}$ of continuous functions $u_n : [a, b] \to \mathbb{R}^N$ that are of uniformly bounded variation

$$\sup_{[a,b]} |u_n| + TV_{[a,b]}\,(u_n) \le C. \tag{1.3}$$

The products $g(u_n)du_n$ are well defined by (1.2) and belong to $\mathcal{M}[a, b] = \big[\mathcal{C}[a, b]\big]^\star$, the dual space of $\mathcal{C}\big([a, b]; \mathbb{R}^N\big)$. The space of Radon measures $\mathcal{M}[a, b]$ is usually equipped either with the strong topology, generated by the dual norm $\|.\|_{\mathcal{M}}$, or with the weak-$\star$ topology. On account of (1.3), the sequence $\{g(u_n)du_n\}$ satisfies $\|g(u_n)du_n\|_{\mathcal{M}} \le C'$. (Throughout $C, C', \dots$ will stand for constants that are independent of n.) Therefore, along a subsequence,

$$g(u_{n'})\,\frac{du_{n'}}{dx} \rightharpoonup \mu \quad \text{ weak-}\star \text{ in } \mathcal{M}[a, b] \tag{1.4}$$

to some measure μ. Example 1.1 illustrates that, even if $u_n(x) \to u(x)$ pointwise, the sequence $\{g(u_n)du_n\}$ may have multiple limit points in the weak-$\star$ topology.

Example 1.1. Let u_0, u_1 be two states in $\mathbb{R}^N$, $x_0 \in (a, b)$ and $\pi : [0, 1] \to \mathbb{R}^N$ be a Lipschitz continuous path satisfying $\pi(0) = u_0$ and $\pi(1) = u_1$. Consider the sequence of functions v_n, defined by

$$v_n(x) := \begin{cases} u_0 & \text{if } x \in [a, x_0 - 1/n], \\ \pi\big(\frac{x-(x_0-1/n)}{2/n}\big) & \text{if } x \in [x_0 - 1/n, x_0 + 1/n], \\ u_1 & \text{if } x \in [x_0 + 1/n, b]. \end{cases} \tag{1.5}$$

As $n \to \infty$, the sequence $\{v_n\}$ converges pointwise,

$$v_n(x) \to v(x) := \begin{cases} u_0 & \text{if } x \in [a, x_0), \\ \pi(1/2) & \text{if } x = x_0, \\ u_1 & \text{if } x \in (x_0, b], \end{cases} \tag{1.6}$$

and a calculation shows that

$$\int_a^b \varphi(x) g(v_n(x)) \frac{dv_n}{dx}\, dx \to \left(\int_0^1 g(\pi(s))\pi'(s) ds \right) \varphi(x_0)$$

for $\varphi \in \mathcal{C}[a, b]$. That is,

$$g(v_n)\frac{dv_n}{dx} \rightharpoonup c(g, \pi)\delta_{x_0} \quad \text{weak-}\star \text{ in } \mathcal{M}[a, b], \tag{1.7}$$

where δ_{x_0} stands for the Dirac measure at x_0 and the scalar $c(g, \pi)$ is given by the formula

$$c(g, \pi) := \int_0^1 g(\pi(s))\pi'(s) ds. \tag{1.8}$$

Therefore, first, the limit points of $\{g(v_n)\frac{dv_n}{dx}\}$ depend on the limiting graph selected by $\{v_n\}$, expressed via the path π. Second, by mixing sequences whose internal structure is described by several distinct paths π_j, it is easy to generate a sequence $\{v_n\}$ which converges pointwise to the (same) limit v, but where $\{g(v_n)\frac{dv_n}{dx}\}$ has multiple weak-$\star$ limit points. There exists a notable exception to these features: If $g = \nabla f$ for some $f : \mathbb{R}^N \to \mathbb{R}$, then $c(g, \pi) = f(u_1) - f(u_0)$ and the weak-$\star$ limit (1.7) is independent of π. □

To characterize the weak-$\star$ limit points of $\{g(u_n)du_n\}$ we follow the approach of Tartar [29], in his representation theory of weak limits via Young measures. Let $\mathcal{C}_0([a, b]\times\mathbb{R}^N)$ be the space of $\mathbb{R}^N$-valued continuous functions $f = f(x, \lambda)$ that tend to zero as $\lambda \in \mathbb{R}^N$ tends to infinity, equipped with the sup-norm, and let $\mathcal{M}([a, b]\times\mathbb{R}^N) = \big[\mathcal{C}_0([a, b] \times \mathbb{R}^N)\big]^\star$ be the dual space of Radon measures on $[a, b] \times \mathbb{R}^N$. Define the Radon measures p_n by

$$\langle p_n, f\rangle := \int_a^b f(x, u_n(x))\, du_n(x) \qquad \text{for } f \in \mathcal{C}_0([a, b] \times \mathbb{R}^N). \tag{1.9}$$

Then (1.3) implies that

$$|\langle p_n, f\rangle| \le \big(TV_{[a,b]}\,(u_n)\big) \sup_{x\in[a,b],|\lambda|\le C} |f(x, \lambda)|$$

and, hence, $\|p_n\|_{\mathcal{M}} \le C$. There exist a subsequence $\{p_{n_k}\}$ and a measure $p \in \mathcal{M}([a, b] \times \mathbb{R}^N)$ such that

$$p_{n_k} \rightharpoonup p \quad \text{weakly-}\star \text{ in } \mathcal{M}([a, b] \times \mathbb{R}^N). \tag{1.10}$$

The question becomes to characterize the weak-$\star$ limit points of the sequence $\{p_n\}$.

The characterization is effected by using the concept of *graph completion* or (as we prefer to call it) *generalized graph*. This concept was introduced by Bressan and

Rampazzo [3] in a context of control problems and turns out to be sufficiently discriminating to capture the limiting graphs of the sequence $\{u_n\}$. Generalized graphs were used by Dal Maso, LeFloch, and Murat [10] and Raymond [27] as intermediate steps in their definitions of nonconservative products.

DEFINITION 1.2 (see [3]). *A generalized graph of u is a map $(X,U):[0,1]\to[a,b]\times\mathbb{R}^N$ such that X, U are Lipschitz continuous and satisfy*

(1) $(X(0),U(0))=(a,u(a))$, $(X(1),U(1))=(b,u(b))$;

(2) *X is increasing: $s_1<s_2$ implies $X(s_1)\le X(s_2)$;*

(3) *given $y\in[a,b]$, there exists $s\in[0,1]$ such that $X(s)=y$, $U(s)=u(y)$.*

Our aim is to reveal the central role of generalized graphs in providing a geometrically motivated definition of nonconservative products. To this end we exploit an equivalence relation on the space of continuous functions, accounting for reparametrizations of graphs, and the associated pseudometric of uniform graph convergence [3]. By definition, a sequence of graphs $\{gr(u_n)\}$ is Cauchy in the sense of graph convergence, if upon reparametrizing its elements $gr(u_n)$ we obtain a Cauchy sequence in the uniform metric. We will show that, given a sequence of *continuous* functions $\{u_n\}$ that is bounded in $BV[a,b]$, generalized graphs emerge as and are in correspondence to limit points of the sequence of graphs of u_n, $\{gr(u_n)\}$, in the pseudometric of uniform graph convergence. Therefore, the terminology "graph completion" is somewhat misleading, in that it suggests that the completion of the graph is effected arbitrarily from the outside. Since such objects emerge as limits of graphs of sequences of continuous functions, we opt for the more pertinent terminology *generalized graph*. Using this notion we prove a representation theorem on the weak-$\star$ limits in (1.4) and (1.10).

THEOREM 1.3. (a) *Let $\{u_n\}$ be a sequence of continuous functions satisfying the uniform bounds* (1.3). *There exists a subsequence $\{u_{n_k}\}$ and a generalized graph (X,U) such that, for any continuous function $g=g(\lambda)$, we have*

$$\int_{[a,b]}\theta(x)g(u_{n_k}(x))\,du_{n_k}(x)\to\langle\mu(g),\theta\rangle\quad\text{for}\quad\theta\in\mathcal{C}[a,b],\tag{1.11}$$

where $\mu:\mathcal{C}_0(\mathbb{R}^N)\to\mathcal{M}[a,b]$ is defined by

$$\langle\mu(g),\theta\rangle=\int_0^1\theta(X(s))g(U(s))\,dU(s).\tag{1.12}$$

(b) *Conversely, given a generalized graph (X,U), let μ be defined by* (1.12). *There exists a sequence of Lipschitz functions $\{u_n\}$, uniformly bounded in BV, such that for any continuous g,*

$$g(u_n)du_n\rightharpoonup\mu(g)\quad\text{weak-}\star\text{ in }\mathcal{M}[a,b].\tag{1.13}$$

The plan of the article is as follows. Section 2 is preliminary, presenting a change of variable formula for Borel–Stieltjes integrals, an equivalence relation accounting for reparametrizations of continuous paths, and the notion of uniform graph convergence. The case of a continuous BV function is also considered; we introduce the arc-length (or canonical) parametrization of the graph of u and use it, in conjunction with the change of variable formula, to explore the ramifications of definition (1.2) for the nonconservative product $g(u)\,du$, with $u\in\mathcal{C}\cap BV$.

In sections 3 and 4 we study properties of sequences of continuous functions $\{u_n\}$ that are bounded in $BV[a,b]$. After presenting the notion of a generalized graph,

we show that, first, generalized graphs arise as limits of subsequences to $\{gr(u_n)\}$ in the pseudometric of graph convergence and, second, that a given generalized graph can always be approximated by a suitable sequence of graphs of continuous functions. The results are summarized in Theorem 3.2 and are put in a metric space framework at the end of section 3.1. Then in section 3.2 we prove a representation theorem.

The representation theorem suggests to define nonconservative products as measures based on generalized graphs. Two definitions, along with associated weak stability theorems, are pursued: In section 4.1, the nonconservative product is defined as a Radon measure (Definition 4.1), while, in section 4.2, it is defined as a signed Borel measure via its distribution function (Definition 4.4). The definitions are equivalent and invariant under reparametrizations of the geometric graph determined by (X, U); i.e., they depend on the equivalence class of the generalized graph (X, U) but not on the specific representative.

In sections 4.3 and 4.4, we compare various definitions of nonconservative products. To assess the issue, it is instructive to keep in mind the analogy to the solution of the Riemann problem for hyperbolic systems. There exist two approaches for solving the Riemann problem: In the first the solution is effected by patching together elementary solutions (shocks, rarefaction waves, and contact discontinuities), while in the second the whole wave fan is visualized to emerge as a single structure in a small parameter (viscosity, relaxation, etc.) limit of a higher-order theory. Accordingly, two viewpoints for defining nonconservative products can be taken: (i) the product is defined in a pointwise fashion by using a predetermined family of paths at points of jump discontinuity, (ii) the product is defined on the whole structure (the generalized graph). The comparison hinges on the relation between generalized graphs and graphs of functions of bounded variation (Propositions 4.7 and 4.8). The emerging definitions are consistent, with each being more adept for a different range of applications. Section 4.4 analyzes several typical examples of nonconservative products.

We complete the article with a study of the Riemann problem for quasi-linear hyperbolic systems. For genuinely nonlinear systems the solution of the Riemann problem is established in LeFloch [14] and Dal Maso, LeFloch, and Murat [10]. The main step is a construction of the shock curves in the nonconservative case, in the spirit of Lax [13]. The present result is based on an entirely different construction process, following the method of self-similar zero-viscosity limits (see Dafermos [9], Tzavaras [30]). It yields a solution for weak waves of the Riemann problem in the class of general strictly hyperbolic systems with no further assumptions (like genuine nonlinearity or finite number of inflection points) on the characteristic fields. The necessary a priori BV estimates are established in the companion articles [18, 19].

2. Preliminary notions.

2.1. Change of variables formula. Throughout, we work in the framework of functions of bounded variation. The total variation of an $\mathbb{R}^N$-valued function u on an interval $[a, b]$ is defined by

$$TV_{[a,b]}(u) := \sup \sum_{i=1}^{n} |u(x_i) - u(x_{i-1})|,$$

where $|\cdot|$ stands for the Euclidean length in $\mathbb{R}^N$ and the supremum is taken over all finite partitions $a = x_0 < x_1 < \cdots < x_n = b$. Let $u : [a, b] \to \mathbb{R}^N$ be a function of bounded variation and let $T_u : [a, b] \to [0, \infty)$ be the total variation function of u,

defined by

$$T_u(x) := TV_{[a,x]}(u) \qquad \text{for } x \in [a,b]. \tag{2.1}$$

The domain of u can be decomposed into two disjoint sets: $\mathcal{C}_u$ the set of points of continuity of u and $\mathcal{S}_u$ the set of points of discontinuity, respectively. The set $\mathcal{S}_u$ is at most countable, and the right and left limits $u(x+)$, $u(x-)$, for $x \in (a,b)$, and $u(a+)$, $u(b-)$ exist and are finite. We use the notation $u(a-) = u(a)$ and $u(b+) = u(b)$. Note that x is a point of (right or left) continuity for u if and only if x is a point of (right or left) continuity for T_u. In the particular case that u is Lipschitz continuous (or even when u is absolutely continuous, $u \in W^{1,1}\big(a,b)\big)$, the total variation function T_u can be computed by the formula

$$T_u(x) = \int_a^x |u'(y)|\, dy. \tag{2.2}$$

If u is of bounded variation and right continuous on (a,b), there exists a unique finite, signed Borel measure μ_u generated by u,

$$u(x) - u(a+) = \mu_u((a,x]) \quad \text{for } x \in (a,b], \qquad u(a+) - u(a) = \mu_u(\{a\}).$$

The measure μ_u is typically denoted by du, its total variation measure satisfies $|du| = dT_u$, and it can be decomposed into an absolutely continuous part $u'(x)dx$, an atomic part $d_a u$, and a singular part (relative to the Lebesgue measure) $d_s u$, according to the formula $du = u'(x)dx + d_a u + d_s u$.

For functions u, $v : [a,b] \to \mathbb{R}^N$ right continuous and of bounded variation, there is an integration by parts formula: If u and v have no common points of discontinuity, $\mathcal{S}_u \cap \mathcal{S}_v = \emptyset$, then

$$\int_{[\alpha,\beta]} v(x)du(x) + \int_{[\alpha,\beta]} u(x)dv(x) = v(\beta+)u(\beta+) - v(\alpha-)u(\alpha-) \tag{2.3}$$

for any $[\alpha,\beta] \subset [a,b]$. (Here and in what follows we use the notation $v\,du$ to mean the inner product $\sum_i v_i du_i$, where u_i and v_i are the components of u and v, respectively.) If v is absolutely continuous, (2.3) takes the more conventional form

$$\int_{[\alpha,\beta]} v(x)du(x) = -\int_{[\alpha,\beta]} u(x)v'(x)dx + v(\beta)u(\beta+) - v(\alpha)u(\alpha-). \tag{2.4}$$

We will need certain change of variable formulas that follow from a general measure theoretic construction. We first outline the general construction of image measures, taken out of Folland [11, p. 287]. Let $(\Omega, \mathcal{B}, \mu)$ be a measure space, let $(\Omega', \mathcal{B}')$ be a measurable space, and let $\varphi : \Omega \to \Omega'$ be a $(\mathcal{B}, \mathcal{B}')$-measurable map. Then μ induces an image measure μ^φ on Ω' by

$$\mu^\varphi(E) = \mu\big(\varphi^{-1}(E)\big) \qquad \text{for } E \in \mathcal{B}'. \tag{2.5}$$

It is easy to check that μ^φ defines a measure on $(\Omega', \mathcal{B}')$. (The reader is warned not to confuse the measure μ^φ with the Borel measure μ_u generated by the right continuous BV function u.) One also has the formula.

PROPOSITION 2.1. *If $f : \Omega' \to \mathbb{R}$ is a measurable function, then*

$$\int_{\Omega'} f\, d\mu^\varphi = \int_\Omega \big(f \circ \varphi\big) d\mu \tag{2.6}$$

whenever either side is defined.

The proof of (2.6) follows the familiar process of first proving it for characteristic functions $f = \mathbb{1}_E$ with $E \in \mathcal{B}'$, by using $\mathbb{1}_E \circ \varphi = \mathbb{1}_{\varphi^{-1}(E)}$ and (2.5), then for simple functions and finally for integrable functions; cf. [11, p. 287]. In probability theory, when μ is a probability measure and $\varphi : \Omega \to \mathbb{R}$ is a Borel-measurable real-valued function, the image measure μ^φ is called the distribution of the random variable φ.

For u a right continuous function of bounded variation, let $L^1(du)$ denote the integrable functions with respect to the (signed) vector measure du. For instance, all the bounded, Borel measurable functions belong to $L^1(du)$. Proposition 2.1 provides certain change of variable formulas for Borel–Stieltjes integrals that are used extensively in the sequel.

THEOREM 2.2. *Let $u : [a,b] \to \mathbb{R}^N$ be a right continuous function of bounded variation, and let $X : [0,1] \to [a,b]$ be a continuous increasing (not necessarily strictly increasing) change of variables with $X(0) = a$, $X(1) = b$.*

(a) *If X^{-1} denotes the left-continuous inverse of X, then, for $f \in L^1\big(d(u \circ X)\big)$, we have*

$$\int_{[0,1]} f(s)\, d\big(u \circ X\big)(s) = \int_{[a,b]} f \circ X^{-1}(x)\, du(x). \tag{2.7}$$

(b) *For any function $g \in L^1\big(du\big)$, we have*

$$\int_{[0,1]} \big(g \circ X\big)(s)\, d\big(u \circ X\big)(s) = \int_{[a,b]} g(x)\, du(x). \tag{2.8}$$

Formula (2.8) when du is the Lebesgue measure is stated as an exercise in Folland [11, p. 103]. It is easy to construct examples showing that (2.8) fails if the hypothesis "X continuous" is replaced by "X right continuous."

Proof. We first establish (2.7) and (2.8) under the hypotheses

$$\begin{aligned} &u : [a,b] \to \mathbb{R} \quad \text{increasing and right continuous,} \\ &f : [0,1] \to [0,\infty] \quad \text{Borel measurable,} \\ &g : [a,b] \to [0,\infty] \quad \text{Borel measurable.} \end{aligned} \tag{2.9}$$

Since X is increasing, the inverse of X is a multivalued increasing map. We select the single-valued left-continuous inverse $\varphi = X^{-1}$ of the map X. Note that $X \circ \varphi = id$, but in general $\varphi \circ X \neq id$. The function $\varphi : [a,b] \to [0,1]$ is single valued, increasing, and satisfies

$$\varphi^{-1}\big((s,\tau]\big) = (X(s), X(\tau)] \quad \text{for } s, \tau \in [0,1].$$

Since the half-open intervals generate the Borel σ-algebra, φ is a $(\mathcal{B}_{[a,b]}, \mathcal{B}_{[0,1]})$-measurable map, that is, a Borel measurable map. Also, $f \circ \varphi$ is Borel measurable as well.

Let μ_u be the Borel measure generated by u, and let μ_u^φ be the image measure of μ_u under φ. Then

$$\mu_u^\varphi\big((s,\tau]\big) = \mu_u\big(\varphi^{-1}\big((s,\tau]\big)\big) = \mu_u\big((X(s), X(\tau)]\big) = \mu_{u\circ X}\big((s,\tau]\big)\,.$$

Since μ_u^φ and $\mu_{u\circ X}$ agree on the half-open intervals, the extension theorems for premeasures (e.g., [11, Thms. 1.14 and 1.16]) imply $\mu_u^\varphi = \mu_{u\circ X}$ on the Borel sets $\mathcal{B}_{[0,1]}$. Formula (2.7) is then a consequence of Proposition 2.1. In turn, (2.8) follows from (2.7), upon setting $f = g \circ X$ and using the identity $X \circ \varphi = id$.

Once (2.7) and (2.8) are established under (2.9), they are extended to hold under the hypotheses of Theorem 2.2. Consider, for instance, (2.7). It is first extended to hold for Borel measurable functions $f : [0,1] \to \mathbb{R}$ that are integrable with respect to $d(u \circ X)$, by using the decomposition $f = f^+ - f^-$, with f^+, $f^- \in L^1\big(d(u \circ X)\big)$. Next, if $u : [a,b] \to \mathbb{R}$ is a function of bounded variation, it can be decomposed in the form $u = u_1 - u_2$, with u_1, u_2 increasing and thus du_1, du_2 positive measures. Using the induced decomposition $d(u \circ X) = d(u_1 \circ X) - d(u_2 \circ X)$ of the signed measure $d(u \circ X)$ into a difference of positive measures, we can extend (2.7) to hold in this case also. Finally, the extension to the vector-valued case is trivial. □

Theorem 2.2 also yields a simple proof of the chain rule for Lipschitz functions in the one-dimensional context (see Marcus and Mizel [21], Boccardo and Murat [2]).

COROLLARY 2.3. *Suppose that $u : [a,b] \to \mathbb{R}^N$ is absolutely continuous and $X : [0,1] \to [a,b]$ is increasing, continuous, and onto. Then*

$$d\big(u \circ X\big) = \big(u' \circ X\big)\, dX. \tag{2.10}$$

If X is absolutely continuous, then

$$\frac{d}{ds}\big(u \circ X\big)(s) = u'\big(X(s)\big)\, \frac{dX}{ds}(s) \qquad \text{for almost everywhere (a.e.)} \quad s \in [0,1]. \tag{2.11}$$

Proof. We will show that

$$\int_0^s d(u \circ X) = \int_0^s u'(X(s))\, dX(s) \qquad \text{for } s \in [0,1].$$

Fix $s \in [0,1]$ and let $y = X(s)$ and $\bar{s} = \inf\{s \in [0,1] : X(s) > y\}$. Then $X(\tau) = y$ on the interval $[s, \bar{s}]$, and (2.8) in Theorem 2.2 implies

$$\begin{aligned}\int_0^s d(u \circ X) = \int_0^{\bar{s}} d(u \circ X) = \int_{[a,y]} du(x) = \int_{[a,y]} u'(x)dx \\ = \int_0^{\bar{s}} u'(X(s))dX(s) = \int_0^s u'(X(s))dX(s).\end{aligned}$$

Hence, (2.10) follows.

If X is absolutely continuous, then $u \circ X$ is also absolutely continuous and (2.11) follows from (2.10). □

Let $BV[a,b]$ be the set of all functions $u : [a,b] \to \mathbb{R}^N$ of bounded variation. The space $BV[a,b]$ can be identified to the space of (equivalence classes of) functions u in $L^1(a,b)$ whose distributional derivative, du/dx, is a finite, signed Borel measure. To see that, let $u \in BV[a,b]$ and let $\bar{u}$ denote a right continuous BV function such that $u = \bar{u}$ a.e. (the function $\bar{u}$ is uniquely determined by the equivalence class of u). By the Riesz representation theorem, the signed Borel measure $d\bar{u}$, generated by $\bar{u}$, can be identified with a bounded linear functional $\nu_{\bar{u}}$ on $\mathcal{C}[a,b]$,

$$\langle \nu_{\bar{u}}, \theta \rangle = \int_{[a,b]} \theta(x) d\bar{u}(x) \qquad \text{for } \theta \in \mathcal{C}[a,b]. \tag{2.12}$$

Then (2.4) implies that, for $\varphi \in \mathcal{C}_c^1(a,b)$,

$$\langle \nu_{\bar{u}}, \varphi \rangle = \int_{(a,b)} \varphi(x) d\bar{u}(x) = -\int_{(a,b)} \varphi'(x) u(x) dx\,, \tag{2.13}$$

i.e., the distributional derivative of u satisfies $du/dx = \nu_{\bar{u}}$. Moreover,

$$|\nu_{\bar{u}}|\big([a,b]\big) = TV_{[a,b]}(\bar{u})\,. \tag{2.14}$$

We note that if another representative is used on the right of (2.14), then equality is in general replaced by a strict inequality. The space $BV[a,b]$, when equipped with the norm

$$\|u\|_{BV} = \|u\|_{L^1} + |\nu_{\bar{u}}|\big([a,b]\big),$$

becomes a Banach space. For functions of one variable, it is customary to use the equivalent norm

$$\|u\|_{BV} = \|u\|_{L^\infty} + |\nu_{\bar{u}}|\big([a,b]\big).$$

We refer to Folland [11] and Volpert [31] for further information on the theory of BV functions.

2.2. Reparametrizations and distance of graphs. We present first the notion of uniform graph convergence [3, 10], which emerges when continuous paths are studied from the viewpoint of identifying two paths if their ranges coincide. In $\mathcal{C}[0,1]$, the space of continuous paths $V : [0,1] \to \mathbb{R}^M$, an equivalence relation is introduced.

DEFINITION 2.4. *We say that V_1 and V_2 are equivalent, $V_1 \sim V_2$, if and only if there exist two continuous, increasing (but not necessarily strictly increasing) and surjective maps γ_1, $\gamma_2 : [0,1] \to [0,1]$ such that $V_1 \circ \gamma_1 = V_2 \circ \gamma_2$.*

The following lemma is proved in [3, Lemma 1].

LEMMA 2.5. *Let V_1, $V_2 \in \mathcal{C}[0,1]$. Given two continuous, increasing, and surjective maps γ_1, $\gamma_2 : [0,1] \to [0,1]$ there exist two increasing, surjective maps α_1, $\alpha_2 : [0,1] \to [0,1]$, Lipschitz continuous with Lipschitz constant 3, such that*

$$\max_{[0,1]} \big|V_1 \circ \alpha_1 - V_2 \circ \alpha_2\big| = \max_{[0,1]} \big|V_1 \circ \gamma_1 - V_2 \circ \gamma_2\big|.$$

Therefore, $V_1 \sim V_2$ if and only if there exist two Lipschitz continuous, increasing, and surjective maps α_1, $\alpha_2 : [0,1] \to [0,1]$ such that $V_1 \circ \alpha_1 = V_2 \circ \alpha_2$.

If V is continuous and of bounded variation (and $TV_{[0,1]}(V) \neq 0$), then $V^c : [0,1] \to \mathbb{R}^M$, the *canonical parametrization* of V, is defined by

$$V^c(\tau) = V(s),\ \tau = \frac{1}{L}\,T_V(s), \qquad \text{where } L := TV_{[0,1]}(V), \tag{2.15}$$

the total variation function T_V being defined by (2.1). It is easy to check that V^c is well defined and, for $\tau_1 < \tau_2$,

$$|V^c(\tau_2) - V^c(\tau_1)| = |V(s_2) - V(s_1)| \le T_V(s_2) - T_V(s_1) = L(\tau_2 - \tau_1). \tag{2.16}$$

Hence, $V = V^c \circ \gamma$ where V^c is a Lipschitz continuous path and $\gamma = (1/L)T_V$ is continuous. The equivalence relation separates $\mathcal{C}[0,1]$ into equivalence classes that satisfy the following properties:

(1) If $V_1 \sim V_2$ and V_1 is of bounded variation, then V_2 is of bounded variation.

(2) If V is of bounded variation, then a Lipschitz continuous representative of the class can be selected, V^c.

(3) If V_1, V_2 are of bounded variation, then $V_1 \sim V_2$ if and only if $V_1^c = V_2^c$.

Statements (1) and (2) are clear. To show (3), suppose that $V_1 \sim V_2$ are of bounded variation and introduce the canonical parametrizations $V_\iota = V_\iota^c \circ \gamma_\iota$, where $\gamma_\iota = (1/L_\iota)T_{V_\iota}$ for $\iota = 1, 2$. Let $\alpha_1, \alpha_2 : [0,1] \to [0,1]$ be Lipschitz continuous, increasing, surjective maps such that $V_1 \circ \alpha_1 = V_2 \circ \alpha_2$. Then $T_{V_1} \circ \alpha_1 = T_{V_2} \circ \alpha_2$, $\gamma_1 \circ \alpha_1 = \gamma_2 \circ \alpha_2$, and thus $V_1^c = V_2^c$.

On the space of continuous paths, we define a distance function:

$$\operatorname{dist}(V_1, V_2) := \inf_{\gamma_1, \gamma_2} \max_{s \in [0,1]} |(V_1 \circ \gamma_1)(s) - (V_2 \circ \gamma_2)(s)|, \tag{2.17}$$

where the infimum is taken over all continuous, increasing, and surjective maps $\gamma_1, \gamma_2 : [0,1] \to [0,1]$. Bressan and Rampazzo [3] introduce the distance and show that it defines a pseudometric,

$$\begin{aligned} &\operatorname{dist}(V_1, V_2) = \operatorname{dist}(V_2, V_1), \\ &\operatorname{dist}(V, V) = 0, \\ &\operatorname{dist}(V_1, V_3) \le \operatorname{dist}(V_1, V_2) + \operatorname{dist}(V_2, V_3), \end{aligned}$$

and that, by virtue of Lemma 2.5, the infimum in (2.15) is attained on two Lipschitz continuous paths α_1, α_2, so that the distance can be computed by

$$\operatorname{dist}(V_1, V_2) = \max_{s \in [0,1]} |(V_1 \circ \alpha_1)(s) - (V_2 \circ \alpha_2)(s)|.$$

In particular, that implies $\operatorname{dist}(V_1, V_2) = 0$ if and only if $V_1 \sim V_2$ and, thus, if the distance is viewed on the quotient space $\mathcal{X} = [\mathcal{C}([0,1]; \mathbb{R}^M)/\sim]$, it induces a metric. (Working with equivalence classes has the disadvantage of being cumbersome and identifying otherwise different functions; we will avoid doing that directly, but it is instructive to keep the structure in mind.) The associated convergence is called uniform graph convergence and is denoted by $V_n \xrightarrow{\mathrm{d}} V$: $\{V_n\}$ converges in graph to V if $\operatorname{dist}(V_n, V) \to 0$. Equivalently, $V_n \xrightarrow{\mathrm{d}} V$ if there exist two Lipschitz continuous, increasing, surjective maps $\alpha_n, \alpha : [0,1] \to [0,1]$ such that

$$\operatorname{dist}(V_n, V) = \max_{[0,1]} |V_n \circ \alpha_n - V \circ \alpha| \to 0 \quad \text{as } n \to \infty.$$

Finally we state a compactness result in Proposition 2.6.

PROPOSITION 2.6. *Let $\{V_n\}$ be a sequence of continuous functions on $[0,1]$ that are of uniformly bounded total variation. There exists a subsequence $\{V_{n_k}\}$ and a Lipschitz continuous representative $V^c : [0,1] \to \mathbb{R}^M$ such that $V_{n_k} \xrightarrow{d} V^c$.*

Proof. Let V_n^c be the canonical representatives of V_n, say, $V_n = V_n^c \circ \gamma_n$. By (2.15)–(2.16), V_n^c are uniformly Lipschitz continuous, with Lipschitz constant equal to the uniform variation bound of the sequence V_n. Since $V_n \sim V_n^c$, Lemma 2.5 implies there exist sequences α_n, β_n of uniformly Lipschitz continuous parametrizations such that $V_n \circ \alpha_n = V_n^c \circ \beta_n$. By the Ascoli–Arzela theorem there exist subsequences $V_{n_k}^c$, α_{n_k}, β_{n_k}, and Lipschitz continuous functions $V^c : [0,1] \to \mathbb{R}^M$ and α, $\beta : [0,1] \to [0,1]$ so that $V_{n_k}^c \to V^c$, $\alpha_{n_k} \to \alpha$, and $\beta_{n_k} \to \beta$ uniformly on $[0,1]$. Then

$$\begin{aligned} V_{n_k} \circ \alpha_{n_k} &= V_{n_k}^c \circ \beta_{n_k} \\ &= (V_{n_k}^c \circ \beta_{n_k} - V_{n_k}^c \circ \beta) + V_{n_k}^c \circ \beta \to V^c \circ \beta \end{aligned}$$

uniformly on $[0,1]$ and, thus, $V_{n_k} \xrightarrow{\mathrm{d}} V^c$. ∎

2.3. Nonconservative products for continuous BV functions. The concepts of canonical parametrization and distance of continuous paths have implications when applied to graphs of continuous functions of bounded variation.

Let $u : [a,b] \to \mathbb{R}^N$ be a continuous function of bounded variation. The graph of u,

$$gr(u) := \big\{(x, u(x)) : x \in [a,b]\big\}, \tag{2.18}$$

is a continuous curve in $\mathbb{R} \times \mathbb{R}^N$. We introduce a canonical representative in the spirit of (2.15) (cf. [10]). Let $\sigma : [a,b] \to [0,1]$ be defined by

$$\sigma(x) := \frac{1}{L}\big(x - a + T_u(x)\big), \quad \text{where } L := b - a + TV_{[a,b]}(u) > 0. \tag{2.19}$$

Then σ is strictly increasing, continuous, and surjective and satisfies $\sigma(a) = 0 < \sigma(x) < 1 = \sigma(b)$ for $x \in (a,b)$. The inverse of σ is a function $X : [0,1] \to [a,b]$, which is strictly increasing, continuous, and surjective. If we set $U := u \circ X$, the function $(X,U) : [0,1] \to [a,b] \times \mathbb{R}^N$ is a representative of the graph of u. Further, if $s_1 < s_2$ in $[0,1]$ and y_1, y_2 their respective images under X, $\sigma(y_1) = s_1$ and $\sigma(y_2) = s_2$, then

$$\begin{aligned} X(s_2) - X(s_1) = y_2 - y_1 \le L(\sigma(y_2) - \sigma(y_1)) = L(s_2 - s_1), \\ |U(s_2) - U(s_1)| = |u(y_2) - u(y_1)| \le T_u(y_2) - T_u(y_1) \le L(s_2 - s_1). \end{aligned} \tag{2.20}$$

Hence, (X,U) is Lipschitz continuous with Lipschitz constant L and will be referred to as the *arc-length parametrization* (or canonical representative) of the graph of $u \in \mathcal{C} \cap BV$.

The terminology "arc-length parametrization" is justified as follows: Since $T_u \circ X = T_{u \circ X} = T_U$, the parametrization (X,U) satisfies

$$s = \sigma(X(s)) = \frac{1}{L}\big(X(s) - a + T_U(s)\big) \tag{2.21}$$

for s in $[0,1]$. Therefore, (2.21) implies

$$\frac{dX}{ds} + \left|\frac{dU}{ds}\right| = L, \tag{2.22}$$

which means that the tangent vector to the curve $(X(s), U(s))$ has constant length equal to L. Strictly speaking, the arc-length parametrization corresponds to $L = 1$ in (2.22). This can be attained by stretching the interval $[0,1]$, but we avoid that here.

The graph of a continuous BV function u may be represented by several continuous, increasing, and surjective parametrizations $(Y,V) : [0,1] \to [a,b] \times \mathbb{R}^N$ with Y increasing. The representative can always be chosen to be a Lipschitz continuous path (X,U) with X strictly increasing. The distance between two graphs represented by (Y,V) and $(\bar{Y},\bar{V})$ is defined by $\operatorname{dist}\big((Y,V),(\bar{Y},\bar{V})\big)$ as in (2.17). The notion of distance and the equivalence relation $\sim$ provide a suitable tool for factoring representatives of the same graph (viewed as a geometric object). In what follows, we use the notation $(Y,V) \sim gr(u)$ to denote the general continuous representative (Y,V) of the graph of u and retain the notation (X,U) for the arc-length parametrization or for the associated notion of generalized graph defined in section 3.1.

The arc-length parametrization (X,U) may be used to express the Borel measure du generated by a continuous function of bounded variation u. Using Theorem 2.2,

for the change of variable $x = X(s)$, we obtain

$$\int_{[a,b]} \theta(x)\, du(x) = \int_0^1 \big(\theta \circ X\big)(s) \frac{dU}{ds}\, ds \quad \text{for } \theta \in \mathcal{C}[a,b]. \tag{2.23}$$

The left side in (2.23) is interpreted as a Borel–Stieltjes integral, while the right side is a Lebesgue integral; the formula is useful for theoretical computations involving the measure du. If (Y,V) is an equivalent continuous representative of $gr(u)$, $(Y,V) \sim (X,U)$, repeated use of Theorem 2.2 implies

$$\int_{[a,b]} \theta(x)\, du(x) = \int_0^1 \big(\theta \circ X\big)(s)\, dU(s) = \int_0^1 \big(\theta \circ Y\big)(s)\, dV(s).$$

That is, the Borel measure du depends on $gr(u)$ but not on the particular representative.

We turn now to the definition of nonconservative products for *continuous* functions of bounded variation. A natural way of defining $\mu = g(u)\frac{du}{dx}$ is as a Borel measure, via (1.2). The definition is invariant under reparametrizations of $gr(u)$ and reads

$$\left\langle g(u)\frac{du}{dx}, \theta \right\rangle = \int_0^1 \big(\theta \circ Y\big)(s) g(V(s))\, dV(s) = \int_0^1 \big(\theta \circ X\big)(s) g(U(s)) \frac{dU}{ds}\, ds,$$

where (X,U) is the arc-length parametrization and $(Y,V) \sim gr(u)$ stands for a general representative of the graph of u. This definition is consistent with the one proposed in section 4 for discontinuous BV functions.

3. Generalized graphs.

3.1. Generalized graphs of BV functions. The graph of a general function $u : [a,b] \to \mathbb{R}^N$ of bounded variation has jumps at the points of discontinuity of u. The notion of generalized graph (or graph completion), introduced by Bressan and Rampazzo [3], is an attempt to fill in the jumps by extending the idea of arc-length (or canonical) parametrization.

DEFINITION 3.1. *A generalized graph of u is a map $(X,U) : [0,1] \to [a,b] \times \mathbb{R}^N$ such that X, U are Lipschitz continuous and satisfy the following conditions:*

(1) $(X(0),U(0)) = (a,u(a))$, $(X(1),U(1)) = (b,u(b))$.

(2) *X is increasing: $s_1 < s_2$ implies $X(s_1) \le X(s_2)$.*

(3) *Given $y \in [a,b]$, there exists $s \in [0,1]$ such that $X(s) = y$, $U(s) = u(y)$.*

The range of (X,U) is a compact, connected set containing the graph of u. Let $\sigma = X^{-1}$ be the set theoretic inverse of X; then $\sigma : [a,b] \to [0,1]$ is a strictly increasing, *multivalued* map. The set $\mathcal{C}_\sigma$ of points of continuity of σ (that is, the point where σ is single valued) is dense in $[a,b]$. The set $\mathcal{S}_\sigma$ of points of discontinuity of σ (that is, the points where σ is truly multivalued) is countable and serves as a counter of the jumps and possible loops attached to the graph of u. In this paper, a point $x \in \mathcal{S}_\sigma$ is called a *point of jump* if $u(x-) \neq u(x+)$ and a *loop* if $u(x-) = u(x+)$.

The domain and range of σ admit the decompositions $[a,b] = \mathcal{C}_\sigma \cup \mathcal{S}_\sigma$ and

$$[0,1] = \sigma(\mathcal{C}_\sigma) \cup \sigma(\mathcal{S}_\sigma) = \Big(\bigcup_{y \in \mathcal{C}_\sigma} \{\sigma(y)\} \Big) \cup \Big(\bigcup_{y \in \mathcal{S}_\sigma} [\sigma(y-), \sigma(y+)] \Big), \tag{3.1}$$

respectively. The function u is recovered by the formula

$$u(y) = U(\sigma(y)) \quad \text{for } y \in \mathcal{C}_\sigma. \tag{3.2}$$

The following theorem indicates that the notion of generalized graph captures the limiting graphs selected by pointwise convergent sequences $\{u_n\}$ of *continuous* functions that are stable in $BV[a,b]$. Part (a) of the theorem below provides an extension (and an alternative proof) of the classical Helly selection principle.

THEOREM 3.2. (a) *Let $\{u_n\}$ be a sequence of continuous functions $u_n : [a,b] \to \mathbb{R}^N$ satisfying the uniform bounds* (1.3) *and let (X_n, U_n) be the arc-length parametrizations of $gr(u_n)$. There exists a subsequence $\{u_{n_k}\}$, a function of bounded variation $u : [a,b] \to \mathbb{R}^N$, and an associated generalized graph (X, U) such that*

(1) $(X_{n_k}, U_{n_k}) \xrightarrow{d} (X, U)$;

(2) $\sigma_{n_k}(y) \to \sigma(y)$, $u_{n_k}(y) \to u(y)$ *for all $y \in \mathcal{C}_\sigma$ and a.e. in $[a,b]$,*

where $\sigma_n = X_n^{-1}$ and $\sigma = X^{-1}$ are the set theoretic inverses of X_n and X, respectively.

(b) *Conversely, given a generalized graph (X,U) associated with a BV function u, there exists a sequence $\{u_n\}$ of Lipschitz continuous functions such that*

(1) $\{u_n\}$ *is uniformly bounded in BV,*

(2) $(Y_n, V_n) \xrightarrow{d} (X, U)$ *for any representative $(Y_n, V_n) \sim gr(u_n)$,*

(3) $u_n(y) \to u(y)$ *for $y \in \mathcal{C}_\sigma$ and a.e. in $[a,b]$.*

The proof is based on the following lemma.

LEMMA 3.3. *Suppose that $(X_n, U_n) : [0,1] \to [a,b] \times \mathbb{R}^N$ satisfy the following conditions:*

(1) X_n *is strictly increasing and surjective,*

(2) (X_n, U_n) *are uniformly Lipschitz continuous,*

(3) $(X_n, U_n) \to (X, U)$ *uniformly on $[0,1]$.*

Let $\sigma_n = X_n^{-1}$, $u_n = U_n \circ X_n^{-1}$. Then (X_n, U_n) is a Lipschitz continuous representative of $gr(u_n)$ and

$$\operatorname{dist}\big((Y_n, V_n), (X, U)\big) \to 0 \quad \text{for any representative } (Y_n, V_n) \sim gr(u_n),$$
$$\sigma_n(y) \to \sigma(y), \quad u_n(y) \to u(y) \quad \text{for all } y \in \mathcal{C}_\sigma.$$

Proof. Since X_n is strictly increasing, the functions $\sigma_n = X_n^{-1} : [a,b] \to [0,1]$ and $u_n = U_n \circ X_n^{-1}$ are well defined and continuous. The couple (X_n, U_n) is a representative of the graph of u_n.

Fix $y \in \mathcal{C}_\sigma$ and let $s = \sigma(y)$, $s_n = \sigma_n(y)$. We can write the chain of identities

$$\begin{aligned} s_n - s = \sigma_n(y) - \sigma(y) &= \sigma\Big(X\big(\sigma_n(y)\big)\Big) - \sigma\Big(X_n\big(\sigma_n(y)\big)\Big) \\ &= \sigma\Big(X(s_n)\Big) - \sigma\Big(X_n(s_n)\Big). \end{aligned}$$

Since $X_n \to X$ uniformly and $X_n(s_n) = y \in \mathcal{C}_\sigma$, we deduce $s_n \to s$.

Next, assumption (2) implies

$$|U_n(s_n) - U_n(s)| \le \operatorname{Lip}(U_n)\,|s_n - s| \to 0.$$

Hence, $u_n(y) = U_n(s_n) \to U(s) = u(y)$.

Finally, if (Y_n, V_n) is any continuous representative of $gr(u_n)$, then

$$\operatorname{dist}\big((Y_n, V_n), (X, U)\big) \le \operatorname{dist}\big((X_n, U_n), (X, U)\big) \to 0 \quad \text{as } n \to \infty.$$

This completes the proof. □

Proof of Theorem 3.2. (a) The sequence $\{u_n\}$ consists of continuous functions. Let (X_n, U_n) be the arc-length parametrizations of $gr(u_n)$, defined by inverting

$$\sigma_n(x) := \frac{1}{L_n}\big(x - a + T_{u_n}(x)\big), \qquad L_n := b - a + T_{u_n}(b),$$

and setting $X_n := {\sigma_n}^{-1}$, $U_n := u_n \circ X_n$. In view of (2.22) and (1.3), (X_n, U_n) are uniformly Lipschitz continuous. There exists a subsequence (X_{n_k}, U_{n_k}) and a Lipschitz continuous function $(X, U) : [0,1] \to [a,b] \times \mathbb{R}^N$ such that

$$X_{n_k} \to X, \quad U_{n_k} \to U \qquad \text{uniformly on } [0,1]. \tag{3.3}$$

Hence, $\operatorname{dist}((X_{n_k}, U_{n_k}), (X, U)) \to 0$ as $k \to \infty$, and (3.3), in conjunction with Lemma 3.3, yields the conclusion of part (a).

(b) Given a generalized graph (X, U), let (X_n, U_n) be defined by

$$X_n := \left(1 - \frac{1}{n}\right) X + \frac{1}{n}\big(a + (b-a)s\big), \qquad U_n := U.$$

Then $X_n : [0,1] \to [a,b]$ is strictly increasing, Lipschitz continuous, and surjective; $\{(X_n, U_n)\}$ are uniformly Lipschitz continuous, while $(X_n, U_n) \to (X, U)$ uniformly on $[0,1]$. The functions u_n, defined by $u_n = U_n \circ X_n^{-1}$, are Lipschitz continuous and satisfy

$$\begin{aligned} \sup_{[a,b]} |u_n| &= \sup_{[0,1]} |U|, \\ TV_{[a,b]}(u_n) &= TV_{[0,1]}(U) \le Lip(U). \end{aligned}$$

The conclusion of part (b) now follows from Lemma 3.3. ∎

It is instructive to place the above concepts in a functional analysis framework. Let

$$E = \big\{(Y,V) \in \mathcal{C}\big([0,1]; [a,b] \times \mathbb{R}^N\big) \ : \ Y(0) = a,\ Y(1) = b\big\} \tag{3.4}$$

and $\mathcal{X} := (E/\sim)$ be the quotient space of E over the equivalence relation $\sim$ introduced in Definition 2.4. The elements of $\mathcal{X}$ are equivalence classes of functions: (Y_1, V_1), (Y_2, V_2) are in the same equivalence class if and only if $(Y_1, V_1) \circ \alpha = (Y_2, V_2) \circ \beta$ for some Lipschitz and increasing reparametrizations of $[0,1]$; that is, the curves determined by the functions (Y_1, V_1) and (Y_2, V_2) coincide. The elements of $\mathcal{X}$ can thus be visualized as geometric curves in $[a,b] \times \mathbb{R}^N$ with $Y(0) = a$, $Y(1) = b$.

If $(Y, V) \in E \cap BV$, one can select, using (2.19), a Lipschitz continuous representative of the equivalence class $[(Y,V)]$. This representative is denoted here by (X, U) and is characteristic to the class. The reason is that, for (Y,V) and $(\bar{Y}, \bar{V})$ of bounded variation, $(Y,V) \sim (\bar{Y}, \bar{V})$ if and only if the corresponding canonical representatives of $[(Y,V)]$ and $[(\bar{Y}, \bar{V})]$ are identical, $(X,U) = (\bar{X}, \bar{U})$. We emphasize that we can talk about the canonical representative only for $\mathcal{C} \cap BV$ curves. (Recall that if one representative of the equivalence class is of bounded variation, any representative is of bounded variation.)

When $\mathcal{X}$ is equipped with the pseudometric $\operatorname{dist}((Y,V), (\bar{Y}, \bar{V}))$, defined in (2.17), it becomes a metric space. Consider now the sets

$$\begin{aligned} \mathcal{F} &= \big\{[(Y,V)] \in \mathcal{X} \ : \ (Y,V) \text{ is of bounded variation and } X \text{ is strictly increasing}\big\}, \\ \mathcal{G} &= \big\{[(Y,V)] \in \mathcal{X} \ : \ (Y,V) \text{ is of bounded variation and } X \text{ is increasing}\big\}, \end{aligned} \tag{3.5}$$

where (X,U) always refers to the canonical representative of $[(Y,V)]$. Section 2.3 indicates that $\mathcal{F}$ can be identified with the set of continuous functions of bounded variation, $(\mathcal{C} \cap BV)\big([a,b]; \mathbb{R}^N\big)$. The elements of $\mathcal{F}$ are viewed as the graphs of the functions

u, with the canonical representative coinciding with the arc-length parametrization of $gr(u)$ in (2.19). The set $\mathcal{G}$ is the closure of $\mathcal{F}$ in the metric induced by the distance function (2.17) and may itself be viewed as a complete metric space. The canonical representative of each equivalence class element of $\mathcal{G}$ is a generalized graph in the sense of Definition 3.1. Henceforth, elements of $\mathcal{G}$ are denoted by $gr(X,U)$ and are visualized as the geometric graphs generated by (X,U). We remark that elements of $\mathcal{G}$ are not in correspondence with the space of BV functions, but rather $\mathcal{G}$ consists of all possible limit points of $\mathcal{F}$ in the distance metric.

3.2. Representation of weak-$\star$ limits. Consider a sequence $\{u_n\}$ of *continuous* functions satisfying the uniform bounds (1.3). The sequence $\{g(u_n)du_n\}$ may have multiple limit points in the weak-$\star$ topology of $\mathcal{M}[a,b]$ (cf. Example 1.1). We now characterize such limits for any continuous g in the following representation theorem.

THEOREM 3.4. (a) *Let $\{u_n\}$ be a sequence of continuous functions satisfying the uniform bounds* (1.3). *There exists a subsequence $\{u_{n_k}\}$ and a generalized graph (X,U) such that, for any continuous function $g=g(\lambda)$, we have*

$$\int_{[a,b]} \theta(x) g(u_{n_k}(x))\, du_{n_k}(x) \to \langle \mu(g), \theta\rangle \quad \text{for} \quad \theta \in \mathcal{C}[a,b], \tag{3.6}$$

where $\mu : \mathcal{C}_0(\mathbb{R}^N) \to \mathcal{M}[a,b]$ is defined by

$$\langle \mu(g), \theta\rangle = \int_0^1 \theta(X(s)) g(U(s))\, dU(s). \tag{3.7}$$

(b) *Conversely, given a generalized graph (X,U), let μ be defined by* (3.7). *There exists a sequence of Lipschitz functions $\{u_n\}$, uniformly bounded in BV, such that for any continuous g,*

$$g(u_n)du_n \rightharpoonup \mu(g) \quad \text{weak-}\star \text{ in } \mathcal{M}[a,b]. \tag{3.8}$$

Theorem 3.4 is based on a characterization of the weak-$\star$ limit points to the sequence of Radon measures $\{p_n\}$ defined in (1.9). The key ingredient is the following weak stability type of theorem.

THEOREM 3.5. *Let $\{u_n\}$ be a sequence of continuous functions $u_n : [a,b] \to \mathbb{R}^N$ satisfying the uniform bounds* (1.3), *and let (X_n,U_n) be the arc-length parametrization of $gr(u_n)$. If*

$$(X_n,U_n) \xrightarrow{d} (X,U) \tag{3.9}$$

to some generalized graph (X,U) associated with a BV function u, then

$$\int_{[a,b]} f(x,u_n(x))\, du_n(x) \to \int_0^1 f(X(s),U(s))dU(s) \quad \text{for } f \in \mathcal{C}_0([a,b]\times\mathbb{R}^N). \tag{3.10}$$

Proof. Let (X_n,U_n) be the arc-length parametrizations of $gr(u_n)$, and let (X,U) be a generalized graph of u. Hypothesis (3.9) implies that for some α_n and α, Lipschitz continuous reparametrizations of the interval $[0,1]$, we have

$$\begin{aligned} Y_n &:= X_n \circ \alpha_n \to X \circ \alpha =: Y \qquad \text{uniformly on } [0,1],\\ V_n &:= U_n \circ \alpha_n \to U \circ \alpha =: V \qquad \text{uniformly on } [0,1]. \end{aligned}$$

By virtue of Theorem 2.2, we may express the integrals

$$\int_{[a,b]} f(x,u_n(x))\,du_n(x) = \int_0^1 f(X_n(s),U_n(s))dU_n(s) = \int_0^1 f(Y_n(s),V_n(s))dV_n(s),$$

$$\int_0^1 f(X(s),U(s))dU(s) = \int_0^1 f(Y(s),V(s))dV(s).$$

Fix $f \in \mathcal{C}_0\big([a,b]\times\mathbb{R}^N\big)$. Note that (Y_n,V_n) and (Y,V) are continuous and satisfy

$$\begin{aligned} &TV(V_n) = TV(U_n) \le C,\\ &(Y_n,V_n) \to (Y,V) \quad \text{uniformly on } [0,1]. \end{aligned} \tag{3.11}$$

It suffices to show that (3.11) implies

$$\int_0^1 f(Y_n(s),V_n(s))dV_n(s) \to \int_0^1 f(Y(s),V(s))dV(s). \tag{3.12}$$

Step 1. We first show that, if V_n, $V : [0,1] \to \mathbb{R}^N$ are functions of bounded variation (not necessarily continuous) such that

(1) $\|V_n - V\|_\infty \to 0$,

(2) $TV(V_n) \le C$,

then, for any $[\alpha,\beta] \subset [0,1]$ and $\varphi \in \mathcal{C}[\alpha,\beta]$, we have

$$\int_{[\alpha,\beta]} \varphi(s)\,dV_n(s) \to \int_{[\alpha,\beta]} \varphi(s)\,dV(s). \tag{3.13}$$

If $\psi \in \mathcal{C}^1[\alpha,\beta]$, then (2.4) implies

$$\begin{aligned} \int_{[\alpha,\beta]} \psi(s)\,dV_n(s) &= -\int_{[\alpha,\beta]} \psi'(s)\,V_n(s)ds + \psi(\beta)V_n(\beta+) - \psi(\alpha)V_n(\alpha-)\\ &\to -\int_{[\alpha,\beta]} \psi'(s)\,V(s)ds + \psi(\beta)V(\beta+) - \psi(\alpha)V(\alpha-)\\ &= \int_{[\alpha,\beta]} \psi(s)\,dV(s). \end{aligned} \tag{3.14}$$

Given $\varphi \in \mathcal{C}[\alpha,\beta]$, there exists for every $\varepsilon > 0$ a function $\psi \in \mathcal{C}^1[\alpha,\beta]$ such that $\|\psi - \varphi\|_\infty < \varepsilon$. The relation

$$\begin{aligned} &\left|\int_{[\alpha,\beta]} \varphi(s)\,dV_n(s) - \int_{[\alpha,\beta]} \varphi(s)\,dV(s)\right|\\ &\le \varepsilon\, TV_{[\alpha,\beta]}(V_n) + \left|\int_{[\alpha,\beta]} \psi(s)\,dV_n(s) - \int_{[\alpha,\beta]} \psi(s)\,dV(s)\right| + \varepsilon\, TV_{[\alpha,\beta]}(V), \end{aligned}$$

in conjunction with (3.14), yields (3.13).

Step 2. Step 1, in conjunction with (3.11), implies

$$\int_0^1 f(Y(s),V(s))dV_n(s) \to \int_0^1 f(Y(s),V(s))dV(s).$$

On the other hand, again by (3.11),

$$\left|\int_0^1 (f(Y_n,V_n) - f(Y,V))\,dV_n\right| \le (\max|f(Y_n,V_n) - f(Y,V)|)\int_{[0,1]} |dV_n| \to 0,$$

as $n \to \infty$. Hence, (3.12) follows. ∎

Proof of Theorem 3.4. (a) Let $\{u_n\}$ be the sequence of continuous functions satisfying (1.3), and let (X_n, U_n) be the arc-length parametrizations of $gr(u_n)$; the latter are uniformly Lipschitz continuous. Let $\{p_n\}$ be the sequence of Radon measures defined in (1.9). The sequence $\{p_n\}$ is bounded, $\|p_n\|_{\mathcal{M}} \le C$, by the uniform BV bound of $\{u_n\}$.

Let p be a weak-$\star$ limit point of $\{p_n\}$. For a subsequence

$$\langle p_{n_k}, f(x,\lambda)\rangle = \int_{[a,b]} f(x, u_{n_k}) du_{n_k} \to \langle p, f(x,\lambda)\rangle \text{ for } f \in \mathcal{C}_0([a,b]\times\mathbb{R}^N).$$

Using part (a) of Theorem 3.2 and passing to a further subsequence $\{u_{n_k}\}$, if necessary, we may assume that there is a generalized graph (X,U), so that the arc-length parametrizations $(X_{n_k}, U_{n_k}) \xrightarrow{\mathrm{d}} (X,U)$. Theorem 3.5 implies

$$\langle p, f(x,\lambda)\rangle = \int_0^1 f(X(s), U(s)) dU(s).$$

Taking $f(x,\lambda) = \theta(x)g(\lambda)$ gives the desired result for $g \in \mathcal{C}_0(\mathbb{R}^N)$ and, due to the uniform sup-norm bound of $\{u_n\}$, for any continuous g.

(b) Given a generalized graph (X,U), let μ be defined by (3.7) and let $\{u_n\}$ be the sequence of Lipschitz functions constructed in the proof of part (b) of Theorem 3.2. Then $\{u_n\}$ are uniformly bounded in BV, $\{(X_n, U_n)\}$ are uniformly Lipschitz continuous, and $(X_n, U_n) \to (X,U)$ uniformly on $[0,1]$. Theorem 3.5 for $f(x,\lambda) = \theta(x)g(\lambda)$ implies (3.8). ∎

4. Definition of nonconservative products.

4.1. Definition as a Radon measure. In view of Theorem 3.4, the definition of nonconservative products should be based on a given generalized graph $(X,U) : [0,1] \to [a,b]\times\mathbb{R}^N$ of the function u of bounded variation. The generalized graph (X,U) determines a geometric object (the graph of u together with paths filling the jumps and possible attached loops), call it $gr(X,U)$. We define $g(u)\frac{du}{dx}$ relative to $gr(X,U)$, first as a Radon measure in this section, and then as a finite Borel measure via its distribution function in section 4.2.

Definition 4.1. *Let $(Y,V) \sim gr(X,U)$ denote the general continuous representative of the graph determined by (X,U). Given a continuous map $g : \mathbb{R}^N \to \mathbb{R}^N$, define $\mu(g)$ by*

$$\begin{aligned}\langle \mu(g), \theta\rangle &= \int_0^1 \theta(Y(s)) g(V(s))\, dV(s)\\ &= \int_0^1 \theta(X(s)) g(U(s)) \frac{dU}{ds}\, ds \quad \text{for } \theta \in \mathcal{C}[a,b].\end{aligned} \tag{4.1}$$

Then $\mu(g) \in \mathcal{M}[a,b]$ is called the nonconservative product of $g(u)$ by $\frac{du}{dx}$ and is denoted by

$$\mu(g) = \Big[g(u)\frac{du}{dx}\Big]_{(X,U)}. \tag{4.2}$$

Remark 4.2. (a) We refer to Dal Maso, LeFloch, and Murat [10] for a slightly weaker definition of nonconservative products and to Raymond [27] for a definition

that is equivalent. Comparisons of the various definitions are carried out in section 4.4. References [10, 27] also contain various weak stability results.

(b) Suppose that (Y,V), $(\bar{Y},\bar{V})$ are two representatives of the same graph, that is, $(Y,V)\sim(\bar{Y},\bar{V})$. Then $(Y,V)\circ\beta=(\bar{Y},\bar{V})\circ\alpha$ for some Lipschitz reparametrizations α,β of $[0,1]$. Theorem 2.2 implies the nonconservative product remains invariant,

$$\int_0^1 \theta(Y(s))g(V(s))\,dV(s)=\int_0^1 \theta(\bar{Y}(s))g(\bar{V}(s))\,d\bar{V}(s). \tag{4.3}$$

The measure introduced in Definition 4.1 thus depends on the equivalence class determined by the generalized graph (X,U), i.e., on $gr(X,U)$ as a geometric object. When a Lipschitz representative, such as (X,U) itself is used, then $\langle\mu(g),\theta\rangle$ may be expressed via the last integral in (4.1).

(c) If μ is viewed as a map $\mu:\mathcal{C}_0(\mathbb{R}^N)\to\mathcal{M}[a,b]$, then μ is linear and bounded. The boundedness follows from the estimate

$$|\langle\mu(g),\theta\rangle|\le\big(TV_{[0,1]}(V)\big)\sup_{|\lambda|\le\max|U|}|g(\lambda)|\sup_{x\in[0,1]}|\theta(x)|, \tag{4.4}$$

which implies $\|\mu(g)\|_{\mathcal{M}}\le(TV_{[0,1]}(V))\,\|g\|_{\mathcal{C}_0}$. □

We state next a weak stability theorem for nonconservative products.

THEOREM 4.3. (i) *Let $\{(X_n,U_n)\}$ and (X,U) be generalized graphs. If*

(1) *$TV(U_n)$ is uniformly bounded,*

(2) *$(X_n,U_n)\xrightarrow{d}(X,U)$,*

then

$$\Big[g(u_n)\frac{du_n}{dx}\Big]_{(X_n,U_n)}\rightharpoonup\Big[g(u)\frac{du}{dx}\Big]_{(X,U)}\qquad\text{weak-}\star\text{ in }\mathcal{M}[a,b]. \tag{4.5}$$

(ii) *Let $\{u_n\}$ be a sequence of continuous functions satisfying (1.3), let (X_n,U_n) be the arc-length parametrizations of $gr(u_n)$, and let (X,U) be a generalized graph. If $(X_n,U_n)\xrightarrow{d}(X,U)$, then*

$$g(u_n)du_n\rightharpoonup\Big[g(u)\frac{du}{dx}\Big]_{(X,U)}\qquad\text{weak-}\star\text{ in }\mathcal{M}[a,b]. \tag{4.6}$$

Proof. Define the graphs determined by (X_n,U_n) and (X,U), and let $(Y_n,V_n)\sim gr(X_n,U_n)$ and $(Y,U)\sim gr(X,U)$ be continuous representatives such that $(Y_n,V_n)\to(Y,V)$ uniformly on $[0,1]$. Moreover, $TV(V_n)=TV(U_n)\le C$. The result follows from (3.10) in Theorem 3.5, together with part (b) of Remark 4.2. □

4.2. Distribution functions. We discuss next the properties of nonconservative products when viewed as signed Borel measures defined via their distribution functions. Recall, for a generalized graph (X,U), the set theoretic inverse $\sigma=X^{-1}:[0,1]\to[a,b]$ is a strictly increasing multivalued map.

THEOREM AND DEFINITION 4.4. *Let $(Y,V)\sim gr(X,U)$ be a representative of the graph determined by (X,U). For $x\in[a,b]$ define*

$$F(x)=\int_0^{Y^{-1}(x+)}g(V(s))\,dV(s)=\int_0^{X^{-1}(x+)}g(U(s))\frac{dU}{ds}\,ds, \tag{4.7}$$
$$F(a-)=0.$$

Then F is a right continuous BV function and generates a signed Borel measure μ, determined by

$$\mu((a,x]) = F(x) - F(a) \quad \text{for } x \in (a,b], \qquad \mu(\{a\}) = F(a). \tag{4.8}$$

Also μ coincides with the nonconservative product $[g(u)\frac{du}{dx}]_{(X,U)}$ in (4.1)–(4.2); that is,

$$\begin{aligned} \langle \mu, \theta \rangle &= \int_{[a,b]} \theta(x)\, dF(x) \\ &= \int_{[0,1]} \theta(Y(s))g(V(s))\, dV(s) = \int_0^1 \theta(X(s))g(U(s))\frac{dU}{ds}\, ds \end{aligned} \tag{4.9}$$

for any $\theta \in \mathcal{C}[a,b]$.

Proof. Consider $(Y,V) \sim gr(X,U)$, a general continuous representative of the graph determined by (X,U). We have the following conditions:

(i) $X : [0,1] \to [a,b]$ is Lipschitz continuous, increasing, and surjective with $X(0) = a$, $X(1) = b$, and $X^{-1}(X(s)) = s$ whenever $X(s) \in \mathcal{C}_{X^{-1}}$.

(ii) $Y : [0,1] \to [a,b]$ is continuous, increasing, and surjective with $Y(0) = a$, $Y(1) = b$, and $Y^{-1}(Y(s)) = s$ whenever $Y(s) \in \mathcal{C}_{Y^{-1}}$.

(iii) $(Y,V) \circ \beta = (X,U) \circ \alpha$ for some $\alpha, \beta : [0,1] \to [0,1]$ increasing, Lipschitz, and surjective reparametrizations.

Let $F : [a,b] \to \mathbb{R}^N$ be defined by

$$F(x) = \int_0^{Y^{-1}(x+)} g(V(s))\, dV(s), \quad F(a-) = 0.$$

Then F is a right continuous BV function and generates a signed Borel measure μ, through (4.8). Note that F satisfies

$$F(Y(s)) = \int_0^s g(V)dV \qquad \text{for } s \in Y^{-1}(\mathcal{C}_{Y^{-1}}). \tag{4.10}$$

Step 1. The definition of the distribution function F depends on the equivalence class of (X,U) but not on the specific representative.

It suffices to define F at points $x \in \mathcal{C}_{Y^{-1}}$ and to extend F so that it is right continuous. If $(Y,V) \sim (X,U)$ are two equivalent representatives of $gr(X,U)$, then $Y \circ \beta = X \circ \alpha$, $V \circ \beta = U \circ \alpha$, and $\mathcal{C}_{(X\circ\alpha)^{-1}} \subset \mathcal{C}_{X^{-1}}$, $\mathcal{C}_{(Y\circ\beta)^{-1}} \subset \mathcal{C}_{Y^{-1}}$. For $x \in \mathcal{C}_{(X\circ\alpha)^{-1}} = \mathcal{C}_{(Y\circ\beta)^{-1}}$, Theorem 2.2 implies

$$\begin{aligned} \big[F(x)\big]_{(X,U)} &:= \int_0^{X^{-1}(x)} g(U)dU = \int_0^{(X\circ\alpha)^{-1}(x)} g(U\circ\alpha)d(U\circ\alpha), \\ \big[F(x)\big]_{(Y,V)} &:= \int_0^{Y^{-1}(x)} g(V)dV = \int_0^{(Y\circ\beta)^{-1}(x)} g(V\circ\beta)d(V\circ\beta). \end{aligned}$$

Since such points are dense in $[a,b]$, any of these formulas generates the same distribution function $[F]_{(X,U)} = [F]_{(Y,V)}$ and we may use any representative for calculating F. This shows (4.7).

Step 2. For $\theta \in \mathcal{C}[a,b]$, we shall show that

$$\int_{[a,b]} \theta(x)\, dF(x) = \int_{[0,1]} \theta(X(s))g(U(s))\frac{dU}{ds} ds. \tag{4.11}$$

(Note that this formula is *not* a direct consequence of Theorem 2.2.)

Fix $\psi \in \mathcal{C}^1[a,b]$. Using (2.4), the change of variables $x = X(s)$, (4.7), (4.10), the property $\dot{X} = 0$ on each interval $[\sigma(y-), \sigma(y+)]$ with $y \in \mathcal{S}_\sigma$, and the chain rule for Lipschitz continuous functions, we obtain

$$\begin{aligned}
\int_{[a,b]} \psi(x)\, dF(x) =& \psi(b)F(b+) - \psi(a)F(a-) - \int_{[a,b]} \psi'(x)F(x)\, dx \\
=& \psi(b)F(b+) - \int_0^1 \psi'(X(s))F(X(s))\, \dot{X}(s)\, ds \\
=& \psi(b)F(b+) - \int_{\sigma(\mathcal{C}_\sigma)} \psi'(X(s))\dot{X}(s) \left(\int_0^s g(U)\frac{dU}{d\tau} d\tau \right) ds \\
& - \sum_{y \in \mathcal{S}_\sigma} \int_{\sigma(y-)}^{\sigma(y+)} \psi'(X(s))\dot{X}(s) \left(\int_0^{\sigma(X(s)+)} g(U)\frac{dU}{d\tau} d\tau \right) ds \\
=& \psi(X(1)) \int_0^1 g(U)\frac{dU}{d\tau} d\tau - \int_0^1 \frac{d}{ds}\big(\psi(X(s))\big) \left(\int_0^s g(U)\frac{dU}{d\tau} d\tau \right) ds;
\end{aligned}$$

thus

$$\int_{[a,b]} \psi(x)\, dF(x) = \int_0^1 \psi(X(s))g(U(s))\frac{dU}{ds}\, ds.$$

Since F is of bounded variation and U is Lipschitz continuous, a density argument yields (4.11). The proof of (4.9) follows from part (b) of Remark 4.2. □

Remark 4.5. In view of (4.7) and (4.8), the nonconservative product μ charges points $x \in \mathcal{S}_{X^{-1}}$ according to

$$\mu\big(\{x\}\big) = F(x+) - F(x-) = \int_{X^{-1}(x-)}^{X^{-1}(x+)} g(U(s))\, dU(s). \qquad \square$$

We state and prove a version of the weak stability theorem by using distribution functions.

THEOREM 4.6. *Suppose $\{u_n\}$ is a sequence of continuous functions satisfying (1.3). Let (X_n, U_n) be the arc-length parametrizations of $gr(u_n)$, let (X, U) be a generalized graph, and define the distribution functions*

$$F_n(x) = \int_a^x g(u_n(y))\, du_n(y), \tag{4.12}$$

and $F(x)$ associated with (X, U) by (4.7). If $(X_n, U_n) \xrightarrow{d} (X, U)$, then

$$F_n(x) \to F(x) \qquad \text{a.e. in } (a,b), \tag{4.13}$$

while μ_n and μ, generated by F_n and F, respectively, satisfy $\mu_n \rightharpoonup \mu$ weak-$\star$ in $\mathcal{M}[a,b]$.

Proof. Let (X_n, U_n) be the arc-length parametrizations of $gr(u_n)$; (X_n, U_n) are uniformly Lipschitz. There exist reparametrizations of the interval $[0,1]$, α_n, and α that are uniformly Lipschitz continuous such that $(\bar{X}_n, \bar{U}_n) = (X_n, U_n) \circ \alpha_n$, $(\bar{X}, \bar{U}) = (X, U) \circ \alpha$ satisfy the following: $(\bar{X}_n, \bar{U}_n)$ are uniformly Lipschitz and $\bar{X}_n \to \bar{X}$, $\bar{U}_n \to \bar{U}$ uniformly on $[0,1]$.

Let $\mathcal{S}_{\bar{X}_n^{-1}}$ and $\mathcal{S}_{\bar{X}^{-1}}$ be the points of discontinuity of $\bar{X}_n^{-1}$ and $\bar{X}^{-1}$, respectively, and set $\mathcal{T} = (\bigcup_n \mathcal{S}_{\bar{X}_n^{-1}}) \cup \mathcal{S}_{\bar{X}^{-1}}$. Then $\mathcal{T}$ is countable, and an argument as in the proof of Lemma 3.3 shows

$$\bar{X}_n^{-1}(x) \to \bar{X}^{-1}(x) \quad \text{for } x \in [a,b] \setminus \mathcal{T}. \tag{4.14}$$

Theorem 2.2, for the change of variables $y = \bar{X}_n(s)$, gives

$$F_n(x) = \int_a^x g(u_n(y))\, du_n(y) = \int_0^{\bar{X}_n^{-1}(x)} g(\bar{U}_n(s)) d\bar{U}_n(s) \quad \text{for } x \in \mathcal{C}_{\bar{X}_n^{-1}}. \tag{4.15}$$

An argument, as in the proof of (3.12), shows that

$$\int_0^{\bar{X}^{-1}(x)} g(\bar{U}_n(s)) d\bar{U}_n(s) \to \int_0^{\bar{X}^{-1}(x)} g(\bar{U}(s)) d\bar{U}(s) \quad \text{for } x \in \mathcal{C}_{\bar{X}^{-1}}. \tag{4.16}$$

In turn, (4.14)–(4.16) and the fact that $\bar{U}_n$ are uniformly Lipschitz imply

$$F_n(x) \to F(x) \quad \text{for } x \in [a,b] \setminus \mathcal{T}. \tag{4.17}$$

The distribution functions F_n and F satisfy the following properties: $F_n(a) = 0$, $F(a-) = 0$,

$$F_n(b) = \int_0^1 g(\bar{U}_n(s)) d\bar{U}_n(s)\, ds \to \int_0^1 g(\bar{U}(s)) d\bar{U}(s) = F(b). \tag{4.18}$$

For any test function $\psi \in \mathcal{C}^1[a,b]$, the integration by parts formula (2.4), in conjunction with (4.17)–(4.18), yields

$$\begin{aligned} \int_{[a,b]} \psi(x) dF_n(x) &= -\int_{[a,b]} \psi'(x) F_n(x)\, dx + \psi(b)F_n(b) - \psi(a)F_n(a) \\ &\to -\int_{[a,b]} \psi'(x) F(x)\, dx + \psi(b)F(b) - \psi(a)F(a-); \end{aligned}$$

hence

$$\int_{[a,b]} \psi(x) dF_n(x) \quad \to \quad \int_{[a,b]} \psi(x) dF(x). \tag{4.19}$$

Since F_n are of uniformly bounded variation, (4.19) and a density argument show that $\mu_n \rightharpoonup \mu$ weak-$\star$ in $\mathcal{M}[a,b]$. □

4.3. Generalized graphs and graphs of BV functions. In this section we examine the relation between a generalized graph (X,U) and the graph of the associated BV function u. First observe that Definition 3.1 directly implies the following proposition.

PROPOSITION 4.7. *Let (X,U) be a generalized graph and let $\sigma = X^{-1}$ be the set theoretic inverse of X. Then the following conditions hold:*

(i) *If $y \in \mathcal{C}_\sigma$, then $u(y) = U(\sigma(y))$.*

(ii) *If $y \in \mathcal{S}_\sigma$, then $X(\tau) = y$ and the function*

$$\phi_y(\tau) := U(\tau), \qquad \tau \in J_y := [\sigma(y-), \sigma(y+)],$$

determines a Lipschitz continuous curve that lies on the hyperplane $\{x = y\}$ *and connects* $(y, u(y-))$ *with* $(y, u(y+))$.

(1) *The Lipschitz path* $\phi_y : J_y \to \mathbb{R}^N$ *is either an arc when* $u(y-) \neq u(y+)$ *or a loop when* $u(y-) = u(y+)$.

(2) *The Lipschitz continuity of* U *implies*

$$\sum_{y\in\mathcal{S}_\sigma} \int_{[\sigma(y-),\sigma(y+)]} \left|\frac{\partial \phi_y}{\partial \tau}\right| d\tau \leq \int_0^1 \left|\frac{dU}{ds}\right| ds < \infty. \tag{4.20}$$

(iii) $\mathcal{C}_\sigma \subset \mathcal{C}_u$ *and* $\mathcal{S}_\sigma \supset \mathcal{S}_u$.

A generalized graph completely determines u and also specifies the paths connecting points of discontinuity and possible loops attached to the graph of u. There is no a priori mechanism, given u, for selecting a particular generalized graph. They may be induced by introducing paths at points of discontinuity in $\mathcal{S}_u$ (using straight lines [31] or families of Lipschitz paths [10]) and by possibly attaching loops at points of removable discontinuity or even at points of continuity in $\mathcal{C}_u$ (cf. the examples pointed out in [10] and the notion of extended graph in [27]). A converse to Proposition 4.7 has been proved by Raymond [27].

PROPOSITION 4.8 (see [27]). *Given a function* $u : [a, b] \to \mathbb{R}^N$ *of bounded variation, a countable set* $\mathcal{T}$, *with* $[a, b] \supset \mathcal{T} \supset \mathcal{S}_u$, *and a family of Lipschitz paths* $\Phi = \{\phi_y\}_{y\in\mathcal{T}}$ *such that*

$$\begin{aligned} &\phi_y : [0, 1] \to \mathbb{R}^N \ \textit{is Lipschitz continuous with} \\ &\phi_y(0) = u(y-),\ \phi_y(1) = u(y+) \quad \textit{for } y \in \mathcal{T}, \end{aligned} \tag{A_1}$$

$$\sum_{y\in\mathcal{T}} \int_0^1 \left|\frac{\partial \phi_y}{\partial \tau}\right| d\tau < \infty, \tag{A_2}$$

there exists a generalized graph (X, U) *associated with the triplet* $\big(u, \mathcal{T}, \Phi\big)$.

The triplet $\big(u, \mathcal{T}, \Phi\big)$ is called extended graph in [27]. Apart from its theoretical interest, the proof of the proposition provides a procedure for constructing examples.

Proof. The construction proceeds in two steps.

Step 1. Construction of a continuous, BV representative $(Y, V) : [0, 1] \to [a, b] \times \mathbb{R}^N$ of the graph determined by $\big(u, \mathcal{T}, \Phi\big)$.

Define $q : [a, b] \to [0, 1]$ by setting $q(b) = 1$ and

$$\begin{aligned} q(x) &:= \frac{1}{Q}\left(x - a + \sum_{y\in\mathcal{T}, y<x} \int_0^1 \left|\frac{\partial \phi_y}{\partial \tau}\right| d\tau\right) \qquad \text{for } x \in [a, b), \\ Q &:= b - a + \sum_{y\in\mathcal{T}} \int_0^1 \left|\frac{\partial \phi_y}{\partial \tau}\right| d\tau. \end{aligned} \tag{4.21}$$

Then q is a strictly increasing left-continuous (but generally discontinuous) function satisfying the properties $\mathcal{C}_q = [a, b] \setminus \mathcal{T}$, $\mathcal{S}_q = \mathcal{T}$,

$$\begin{aligned} q(y+) - q(y-) &= \frac{1}{Q} \int_0^1 \left|\frac{\partial \phi_y}{\partial \tau}\right| d\tau, \\ 0 < x_2 - x_1 &\leq Q\big(q(x_2) - q(x_1)\big) \qquad \text{for } x_1 < x_2, \end{aligned} \tag{4.22}$$

and $q(a) = 0 < q(x) < 1 = q(b)$ for $x \in (a, b)$.

The domain and range of q admit the decompositions

$$[a,b] = \mathcal{C}_q \bigcup \mathcal{S}_q,$$
$$[0,1] = q(\mathcal{C}_q) \bigcup \Big(\bigcup_{y\in\mathcal{T}} [q(y-), q(y+)] \Big),$$

where $q(\mathcal{C}_q)$ and each $J_y = [q(y-), q(y+)]$ are mutually disjoint. The closure of the set $q(\mathcal{C}_q)$ is

$$\overline{q(\mathcal{C}_q)} = \Big(\bigcup_{y\in\mathcal{C}_q} \{q(y)\} \Big) \bigcup \Big(\bigcup_{y\in\mathcal{T}} \{q(y-), q(y+)\} \Big).$$

Define now the function (Y, V) as follows:
(a) On each interval $J_y = [q(y-), q(y+)]$ with $y \in \mathcal{T}$, set

$$\begin{cases} Y(s) = y & \text{for } s \in J_y, \\ V(s) = \phi_y\Big(\frac{s-q(y-)}{q(y+)-q(y-)}\Big) & \text{for } s \in J_y. \end{cases} \tag{4.23}$$

(b) On the complement $[0,1] - \cup_{y\in T} J_y = q(\mathcal{C}_q)$, we have $s \in q(\mathcal{C}_q)$ if and only if $s = q(y)$ for precisely one $y \in \mathcal{C}_q$. We define

$$\begin{cases} Y(s) = y & \text{for } s \in q(\mathcal{C}_q), \\ V(s) = u(y) & \text{for } s \in q(\mathcal{C}_q). \end{cases} \tag{4.24}$$

Clearly, Y is an increasing function, (Y, V) are continuous on the interior of each interval J_y, and also for any $s_1, s_2 \in J_y$ with $s_1 < s_2$, we have

$$|V(s_2) - V(s_1)| \le \int_{\frac{s_1-q(y-)}{q(y+)-q(y-)}}^{\frac{s_2-q(y-)}{q(y+)-q(y-)}} \left| \frac{\partial \phi_y}{\partial \tau} \right| d\tau \le \int_0^1 \left| \frac{\partial \phi_y}{\partial \tau} \right| d\tau. \tag{4.25}$$

We proceed to show (Y, V) is continuous for each $s \in \overline{q(\mathcal{C}_q)}$. This follows by a case analysis:

(i) $s \in q(\mathcal{C}_q)$, $s_n \to s$ with $\{s_n\} \subset q(\mathcal{C}_q)$. Then $s_n = q(y_n)$, $s = q(y)$ for some $y_n, y \in \mathcal{C}_q$. By (4.22), $y_n \to y$ and thus

$$Y(s_n) = y_n \to y = Y(s), \quad V(s_n) = u(y_n) \to u(y) = V(s).$$

(ii) $s = q(y-)$ for some $y \in \mathcal{T}$, $s_n \to s$ with $\{s_n\} \subset q(\mathcal{C}_q)$. In this case for large n it is $s_n < s$, and the corresponding points $y_n \in \mathcal{C}_q$ satisfy $s_n = q(y_n)$ and $y_n < y$. Again (4.22) implies

$$0 < y - y_n < Q(q(y-) - q(y_n))$$

and thus, by (A_1),

$$Y(s_n) = y_n \to y- = Y(s), \quad V(s_n) = u(y_n) \to u(y-) = \phi_y(0) = V(s).$$

(iii) If $s = q(y+)$ for some $y \in \mathcal{T}$, $s_n \to s$ with $\{s_n\} \subset q(\mathcal{C}_q)$. Then, as in (ii),

$$Y(s_n) \to y+ = Y(s), \quad V(s_n) = u(y_n) \to u(y+) = \phi_y(1) = V(s).$$

(iv) Now let $s_n \to s$ with $\{s_n\} \subset \overline{q(\mathcal{C}_q)}$. For each n, let $\sigma_n \in q(\mathcal{C}_q)$ such that

$$|\sigma_n - s_n| < \frac{1}{n}, \quad |Y(\sigma_n) - Y(s_n)| < \frac{1}{n}, \quad |V(\sigma_n) - V(s_n)| < \frac{1}{n}.$$

Then $\sigma_n \to s$ and (i)–(iii) imply $Y(s_n) \to Y(s)$, $V(s_n) \to V(s)$.

(v) On the other extreme, let $s_n \to s$ with $\{s_n\} \subset \cup_{y\in\mathcal{T}} J_y$. Then $Y(s_n) = y_n$ with $y_n \in \mathcal{T}$. To simplify the exposition, consider the case $s_n < s$, $s_n \to s$. For each n, define $\sigma_n = q(y_n+)$. Then $\{\sigma_n\} \subset \overline{q(\mathcal{C}_q)}$, $\sigma_n \to s$ and (4.25) implies

$$\sum_n |V(\sigma_n) - V(s_n)| \le \sum_n \int_0^1 \left|\frac{\partial \phi_{y_n}}{\partial \tau}\right| d\tau.$$

It follows from (iv) and hypothesis (A_2) that $Y(s_n) = Y(\sigma_n) \to Y(s)$, $V(s_n) \to V(s)$.

(vi) For general sequences $s_n \to s$, the result follows by combining (iv) and (v).

The function Y is increasing and thus of bounded variation. The total variation of V may be explicitly computed

$$TV_{[0,1]}(V) \le TV_{[a,b]}(u) + \sum_{y\in\mathcal{T}} \left(\int_0^1 \left|\frac{\partial \phi_{y_n}}{\partial \tau}\right| d\tau - \big|u(y+) - u(y-)\big| \right).$$

Thus V is also of bounded variation.

Step 2. Construction of a Lipschitz continuous representative $(X, U) : [0, 1] \to [a, b] \times \mathbb{R}^N$ of the graph determined by $\big(u, \mathcal{T}, \Phi\big)$.

Using the reparametrization (2.15) and the analysis of section 2.2, we can construct the canonical representative of the curve (Y, V). This representative (X, U) is Lipschitz (with Lipschitz constant L) and satisfies

$$(Y, V) = (X, U) \circ \gamma, \text{where } \gamma(s) = \frac{1}{L} T_{(Y,V)}(s),\ s \in [0, 1], \text{ and } L = TV_{[0,1]}(Y, V)\ .$$

Also, X is increasing and (X, U) is a generalized graph. ∎

4.4. Comparison with definitions based on families of paths. In this section, we compare definitions based on families of paths with Theorem 3.4 based on generalized graphs.

We review the definition proposed by Dal Maso, LeFloch, and Murat [10]. This theory is based on a given family of Lipschitz continuous paths $\phi : [0, 1] \times \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}^N$ that satisfy, for some $K > 0$ and for all u_0, $u_1 \in \mathbb{R}^N$ and $\tau \in [0, 1]$, the properties

$$\phi(0; u_0, u_1) = u_0, \quad \phi(1; u_0, u_1) = u_1, \tag{H1}$$

$$\phi(\tau; u_0, u_0) = u_0, \tag{H2}$$

$$\left|\frac{\partial \phi}{\partial \tau}(\tau; u_0, u_1)\right| \le K\,|u_0 - u_1|. \tag{H3}$$

THEOREM AND DEFINITION 4.9 (see [10]). *Let $u : (a, b) \to \mathbb{R}^N$ be a function of bounded variation and $g : \mathbb{R}^N \to \mathbb{R}^N$ be a continuous map. There exists a unique finite signed Borel measure μ on (a, b) such that*

(1) *if u is continuous on a Borel set $B \subset (a, b)$, then*

$$\mu(B) = \int_B g(u)\, du; \tag{4.26}$$

(2) *if u is discontinuous at a point $x \in (a,b)$, then*

$$\mu(\{x\}) = \int_0^1 g(\phi(\tau; u_-, u_+)) \frac{\partial \phi}{\partial \tau}(\tau; u_-, u_+) d\tau \qquad \text{with } u_\pm := u(x\pm). \tag{4.27}$$

The measure μ is called the nonconservative product of $g(u)$ by $\frac{du}{dx}$ and is denoted by

$$\mu = \left[g(u) \frac{du}{dx} \right]_\phi. \tag{4.28}$$

Remark 4.10. The stronger condition

$$\left| \frac{\partial \phi}{\partial \tau}(\tau; u_0, u_1) - \frac{\partial \phi}{\partial \tau}(\tau; v_0, v_1) \right| \le K \left| (u_0 - v_0) - (u_1 - v_1) \right| \tag{H3$'$}$$

is assumed in [10] in place of (H3), in connection with defining products of the form $g(u)\frac{dv}{dx}$, where u and v are BV functions. For instance, (H3$'$) guarantees that such products depend solely on the measure $\frac{dv}{dx}$ and not on the function v. It is straightforward to check that hypotheses (H1)–(H3) suffice for Definition 4.9, for most results presented in [10], and, in particular, for the theorem on weak stability.

It can be checked that the nonconservative product is independent of reparametrizations of the paths and that the definition is consistent with the usual distributional definition in the case of conservative products: if $f : \mathbb{R}^N \to \mathbb{R}$ is a continuously differentiable function, then

$$\left[(Df)(u) \frac{du}{dx} \right]_\phi = \frac{d}{dx}(f(u)). \tag{4.29}$$

The left-hand side in (4.29) is understood in the sense of Definition 4.9, while the right-hand side is understood in the sense of distributions.

Example 4.11. A simple example of paths is the *family of straight lines* ϕ_S, defined by

$$\phi_S(\tau; u_0, u_1) = u_0 + \tau(u_1 - u_0). \tag{4.30}$$

Then (4.27) reads

$$\left[g(u) \frac{du}{dx} \right]_S (\{x\}) = \left(\int_0^1 g(u_- + \tau(u_+ - u_-))\, d\tau \right) (u_+ - u_-)\,,$$

and the nonconservative product coincides with a product introduced by Volpert [31]. To see that, recall that the *averaged superposition* of a BV function $u : (a,b) \to \mathbb{R}^N$ by a continuous function g is the function $\hat{g}(u)$, defined for all $x \in (a,b)$ by

$$\hat{g}(u)(x) = \int_0^1 g(u_- + s(u_+ - u_-))\, ds, \qquad u_\pm := u(x\pm). \tag{4.31}$$

Of course, we have

$$\hat{g}(u)(x) = g(u(x)) \quad \text{for all } x \in \mathcal{C}_u.$$

The function $\hat{g}(u)$ is Borel measurable, and the product $\hat{g}(u)\frac{du}{dx}$ is well defined as a signed Borel measure. This nonconservative product coincides with the one in Definition 4.9 if one uses the family of straight lines:

$$\Big[g(u)\frac{du}{dx}\Big]_S = \hat{g}(u)\frac{du}{dx} \tag{4.32}$$

as Borel measures on (a,b).

A comparison of (4.20) with the hypotheses of Theorem 4.9 indicates that (H2) and (H3) are somewhat restrictive, ruling out the possibility of loops attached to the graph of the BV function u. The gap between the two definitions has been bridged in a definition given by Raymond [27]. It is proved in [27] that this definition is equivalent to Definition 4.1.

In practice, there should be no confusion between the notation introduced in Definitions 4.1 and 4.9, respectively, in view of the following result.

THEOREM 4.12. *Let $u : (a,b) \to \mathbb{R}^N$ be a function of bounded variation and $(X,U) : [0,1] \to [a,b]\times\mathbb{R}^N$ be a generalized graph of u. Suppose there exists a family of paths satisfying* (H1)–(H2) *such that, for every point of discontinuity $x \in \mathcal{S}_u$,*

$$\begin{aligned} &\phi(\tau; u_-, u+) := U\big(s_- + \tau(s_+ - s_-)\big), \qquad \tau\in[0,1],\\ &\text{where } s_\pm := X^{-1}(x\pm),\ \ u_\pm := u(x\pm), \end{aligned} \tag{4.33}$$

and satisfying the "no loop" condition

$$\text{for every } x\in\mathcal{C}_u,\ \text{there exists a unique } s\in[0,1] \text{ such that } X(s)=x. \tag{4.34}$$

Then the nonconservative products in Definitions 4.1 *and* 4.9*, respectively, coincide*

$$\Big[g(u)\frac{du}{dx}\Big]_\phi = \Big[g(u)\frac{du}{dx}\Big]_{(X,U)}$$

as Borel measures on (a,b).

Proof. It will be convenient to view the product in Definition 4.9 as a Borel–Stieltjes integral. Namely, by modifying $g(u)$ at most countably many points, we can construct a function $\overline{g(u)} : [a,b] \to \mathbb{R}^N$ satisfying

$$\overline{g(u)}(x) = g(u(x)) \quad \text{ for } x\in\mathcal{C}_u \tag{4.35}$$

and for $x\in\mathcal{S}_u$

$$\overline{g(u)}\cdot(u(x+)-u(x-)) = \int_0^1 g\big(\phi(\tau;u(x-),u(x+))\big)\partial_\tau\phi(\tau;u(x-),u(x+))\,d\tau. \tag{4.36}$$

Note that the value of $\overline{g(u)}(x)$ is not uniquely determined by points $x\in\mathcal{S}_u$, since any vector othogonal to the jump $u(x+)-u(x-)$ may be added to $\overline{g(u)}(x)$. From Definition 4.9, one deduces that

$$\Big[g(u)\frac{du}{dx}\Big]_\phi = \overline{g(u)}\,du \tag{4.37}$$

as Borel measures on (a,b), where the right-hand side is understood as a Borel–Stieltjes integral. Using the change of variable formula in Theorem 2.2, we thus have

$$\int_{[a,b]}\theta\,\Big[g(u)\frac{du}{dx}\Big]_\phi = \int_{[a,b]}\theta\,\overline{g(u)}\,du = \int_{[0,1]}(\theta\circ X)\,(\overline{g(u)}\circ X)\,d(u\circ X). \tag{4.38}$$

Consider the decomposition $\big([0,1]\setminus J_u\big)\cup J_u$ where $J_u:=\bigcup_x[s_-,s_+]$ with $s_\pm:=X^{-1}(x\pm)$. On one hand, by (4.35) one has $\overline{g(u)}(x)=g(u(x))$ for $x\in\mathcal{C}_u$ and thus, on the set $[0,1]\setminus J_u$, we obtain $u\circ X=U$ and $\overline{g(u)}\circ X=g(u\circ X)=g(U)$. Thus

$$\begin{aligned}\int_{[0,1]\setminus J_u}(\theta\circ X)\,(\overline{g(u)}\circ X)\,d(u\circ X)&=\int_{[0,1]\setminus J_u}(\theta\circ X)\,(g(u)\circ X)\,d(u\circ X)\\&=\int_{[0,1]\setminus J_u}(\theta\circ X)\,g(U)\,\frac{dU}{ds}\,ds.\end{aligned}\tag{4.39}$$

On the other hand, in view of condition (4.34), each interval $[s_-,s_+]\subset J_u$ corresponds to a jump in u, say, $u_\pm:=u(x\pm)$ for some $x\in\mathcal{S}_u$. Using (4.36) and (4.33), we obtain

$$\begin{aligned}\int_{[s_-,s_+]}(\theta\circ X)\,(\overline{g(u)}\circ X)\,d(u\circ X)&=\int_{[s_-,s_+]}\theta(x)\,\overline{g(u)}(x)\,d(u\circ X)\\&=\theta(x)\,\overline{g(u)}(x)\cdot\big(u_+-u_-\big)\\&=\theta(x)\int_0^1 g\big(\phi(\tau;u(x-),u(x+))\big)\,\partial_\tau\phi(\tau;u(x-),u(x+))\,d\tau\\&=\theta(x)\int_{[s_-,s_+]}g(U(s))\,\frac{dU}{ds}\,ds.\end{aligned}\tag{4.40}$$

Combining (4.38)–(4.40) we deduce that

$$\begin{aligned}\int_{[a,b]}\theta\Big[g(u)\frac{du}{dx}\Big]_\phi&=\int_{[0,1]}(\theta\circ X)\,(\overline{g(u)}\circ X)\,d(u\circ X)\\&=\int_{[0,1]}(\theta\circ X)\,g(U)\,\frac{dU}{ds}\,ds\\&=\int_{[a,b]}\theta\Big[g(u)\frac{du}{dx}\Big]_{(X,U)}\end{aligned}$$

for every test function θ. □

Next, we list examples in order to illustrate the relation between regularized sequences $\{v_n\}$, subject to (1.3), and the associated nonconservative products. Since all definitions of nonconservative products are equivalent within their range of applicability, we will use interchangeably the notation $[g(u)\frac{du}{dx}]_\phi$ and $[g(u)\frac{du}{dx}]_{(X,U)}$; the former is applicable when we are given a family of paths ϕ or an extended graph and the latter when we are given a generalized graph (X,U). In any case one can pass from ϕ to (X,U) and vice versa by using Propositions 4.7 and 4.8. We recall that, given a generalized graph (X,U), it is always possible to construct a sequence of smooth functions $\{v_n\}$ that approach (X,U) in the graph distance (cf. Theorem 3.2).

Example 4.13. We return to the sequence $\{v_n\}$ discussed in Example 1.1. With u_0, $u_1\in\mathbb{R}^N$, the functions $\{v_n\}$, v, and the path π defined in (1.5)–(1.6), we have

$$g(v_n)\frac{dv_n}{dx}\rightharpoonup\left(\int_0^1 g(\pi(s))\pi'(s)ds\right)\delta_{x_0}\quad\text{weak-}\star\text{ in }\mathcal{M}[a,b].\tag{4.41}$$

We select the family of paths ϕ so that $\phi(.;u_0,u_1)=\pi$ holds. Then the nonconservative product reads

$$\Big[g(v)\frac{dv}{dx}\Big]_\phi=c(g,\pi)\,\delta_{x_0},\qquad\text{where } c(g,\pi)=\int_0^1 g(\pi(s))\pi'(s)ds\ ,\tag{4.42}$$

and for any g continuous we have $g(v_n)\frac{dv_n}{dx} \rightharpoonup [g(v)\frac{dv}{dx}]_\phi$ weak-$\star$. This example illustrates the important fact that the path ϕ must be selected in agreement with the regularization under consideration. Different regularizations may give rise to different paths ϕ.

When $u_0 \neq u_1$, (4.42) may be interpreted in terms of Definition 4.9. When $u_0 = u_1$, this is no longer possible, because hypothesis (H_2) excludes the possibility of loops. However, it can be interpreted in terms of the more general Definition 4.1 as follows: By Proposition 4.8 the function v together with the location of the loop discontinuity x_0 and the path π determine a generalized graph. A representative (Y, V) of this graph is given by the formulas

(4.43)
$$Y(s) = \begin{cases} a + 3s(x_0 - a), & s \in [0, \frac{1}{3}], \\ x_0, & s \in [\frac{1}{3}, \frac{2}{3}], \\ x_0 + (3s-2)(b - x_0), & s \in [\frac{2}{3}, 1], \end{cases} \qquad V(s) = \begin{cases} u_0, & s \in [0, \frac{1}{3}], \\ \pi(3\,s - 1), & s \in [\frac{1}{3}, \frac{2}{3}], \\ u_1, & s \in [\frac{2}{3}, 1]. \end{cases}$$

Then Definition 4.1 gives

$$\Big[g(v)\frac{dv}{dx}\Big]_{(Y,V)} = c(g;\pi)\,\delta_{x_0} \tag{4.44}$$

and so, as $n \to \infty$,

$$g(v_n)\frac{dv_n}{dx} \rightharpoonup \Big[g(v)\frac{dv}{dx}\Big]_{(Y,V)} \qquad \text{weakly-}\star \in \mathcal{M}[a,b]. \tag{4.45}$$

Note that (4.44) holds for arbitrary u_0 and u_1 and that, when $u_0 = u_1$, the limiting graph (Y, V) contains a loop at the location x_0. □

Example 4.14. Consider next a piecewise constant function $v : [a, b] \to \mathbb{R}^N$ having three points of discontinuity:

$$v(x) = \begin{cases} u_0 & \text{for } x \in [a, c_1), \\ u_1 & \text{for } x \in [c_1, c_2), \\ u_2 & \text{for } x \in [c_2, c_3), \\ u_3 & \text{for } x \in [c_3, b], \end{cases} \tag{4.46}$$

where $a < c_1 < c_2 < c_3 < b$ are real constants, and the u_j's are constant vectors. Let π_j be Lipschitz continuous paths such that $\pi_j(0) = u_{j-1}$ and $\pi_j(1) = u_j$, $j = 1, 2, 3$. In a fashion similar to Example 1.1, we can define a sequence of smooth functions v_n by replacing the jumps in v with smooth transition layers based on the paths π_j such that $\{v_n\}$ are uniformly bounded and $v_n \to v$ pointwise. Then

$$g(v_n)\frac{dv_n}{dx} \rightharpoonup \sum_{j=1,2,3} c(g, \pi_j)\,\delta_{c_j}\,, \quad \text{where} \quad c(g, \pi_j) = \int_0^1 g\big(\pi_j\big)\frac{\partial \pi_j}{\partial s}\,ds\,. \tag{4.47}$$

Accordingly, the nonconservative product is defined so that

$$\Big[g(v)\frac{dv}{dx}\Big]_{(X,U)} = \Big[g(v)\frac{dv}{dx}\Big]_\phi = \sum_{j=1,2,3} c(g, \pi_j)\,\delta_{c_j}. \tag{4.48}$$

In most cases this is done by using Definition 4.9, upon selecting the family of paths ϕ so that $\phi(.; u_j, u_{j+1}) = \pi_j$. There are a few interesting exceptions when one needs

to use Definition 4.1. One is the case where the approximating sequence contains loops. This is discussed in the previous example. Another case is when the jumps of v at $x = c_1$ and $x = c_3$ coincide, $u_0 = u_2$ and $u_1 = u_3$. Then Definition 4.9 prevents us from using, in v_n, different paths for approximating the same jump at two different locations. This difficulty does not arise with Definition 4.1. Upon constructing a representative of the limiting graph, as in the previous example, we define the nonconservative product as in (4.48). ∎

Example 4.15. Given an increasing sequence of points $c_k \in [a,b)$, $k = 0, 1, 2, \dots$, with $c_0 = a$ and $c_\infty := \lim_{k\to\infty} c_k \in (a,b)$, we consider the saltus function $u : [a,b] \to \mathbb{R}^N$ defined by

$$v(x) = \begin{cases} u_0 & \text{for } x \in [a, c_1), \\ u_k & \text{for } x \in [c_k, c_{k+1}), \quad k = 1, 2, \dots, \\ u_\infty & \text{for } x \in [c_\infty, b], \end{cases} \tag{4.49}$$

for constants u_k and u_∞ in $\mathbb{R}^N$. For each jump connecting u_k to u_{k+1} in v, we consider a Lipschitz continuous path $\pi_k(s)$ for $s \in [0,1]$ satisfying $\pi_k(0) = u_{k-1}$ and $\pi_k(1) = u_k$.

Let $c_k^{\pm,n}$, for $k, n = 1, 2, \dots$, be a sequence of points in the interval (a,b) such that $c_k^{-,n} < c_k < c_k^{+,n} < c_{k+1}^{-,n}$, and $c_k^{\pm,n} \to c_k$ as $n \to \infty$. We construct the sequence of regularized functions $v_n : [a,b] \to \mathbb{R}^N$ by

$$v_n(x) = \begin{cases} u_0 & \text{for } x \in [a, c_1^{-,n}), \\ \pi_k\big(\frac{x - c_k^{-,n}}{c_k^{+,n} - c_k^{-,n}}\big) & \text{for } x \in [c_k^{-,n}, c_k^{+,n}), \quad k = 1, 2, \dots, \\ u_k & \text{for } x \in [c_k^{+,n}, c_{k+1}^{-,n}), \quad k = 1, 2, \dots, \\ u_\infty & \text{for } x \in [c_\infty, b]. \end{cases} \tag{4.50}$$

The functions v_n are continuous and

$$TV(v_n) = \sum_{k=1}^{\infty} \int_0^1 \left| \frac{d\pi_k}{ds}(s) \right| ds. \tag{4.51}$$

We assume that the right-hand side of (4.51) is finite, so that the sequence $\{v_n\}$ is of uniformly bounded variation. A calculation shows that

$$g(v_n) \frac{dv_n}{dx} \rightharpoonup \sum_{j=1}^{\infty} c(g, \pi_j)\, \delta_{c_j}, \qquad \text{weak-}\star \text{ in } \mathcal{M}[a,b], \tag{4.52}$$

which suggests that the nonconservative product $[g(v)\frac{dv}{dx}]_{(X,U)}$ should be defined by

$$\Big[g(v) \frac{dv}{dx} \Big]_{(X,U)} = \sum_{k=1}^{\infty} c(g, \pi_k)\, \delta_{c_k}. \tag{4.53}$$

Because of the uniform constant K in hypothesis (H_3) this cannot be handled in general by Definition 4.9. By contrast, Definition 4.1 is adequate to define the nonconservative product as in (4.53); this follows from (A_1)–(A_2) and the construction process in the proof of Proposition 4.8. ∎

5. The Riemann problem for nonconservative hyperbolic systems. The theory developed in the previous sections is now applied to the Riemann problem for first-order quasi-linear hyperbolic systems

$$\partial_t u + A(u)\,\partial_x u = 0, \qquad x \in \mathbb{R},\ t > 0\,, \tag{5.1}$$
$$u(x,0) = \begin{cases} u_- & \text{for } x < 0, \\ u_+ & \text{for } x > 0, \end{cases}$$

where the $N \times N$ matrix $A(u)$ is a smooth function of u, and u_+ and u_- are given vectors in $\mathbb{R}^N$. Because of the invariance of the Riemann problem under dilations $(x,t) \to (\alpha x, \alpha t)$, for $\alpha > 0$, the solution is expected to be a self-similar function of the variable $\xi = x/t$. Accordingly, $u = u(x/t)$ is sought by solving the boundary value problem

$$-\xi \frac{du}{d\xi} + A(u)\,\frac{du}{d\xi} = 0, \tag{P_0}$$
$$u(\pm\infty) = u_\pm\,.$$

In the nonconservative case $A(u)$ is not a Jacobian matrix, and one is confronted with the difficulty of giving an appropriate meaning to the product $A(u)\,du/d\xi$. To address this difficulty, we construct solutions of (P_0) as $\varepsilon \searrow 0$ limits of solutions to

$$-\xi \frac{du_\varepsilon}{d\xi} + A(u_\varepsilon)\,\frac{du_\varepsilon}{d\xi} = \varepsilon \frac{d}{d\xi}\left(B(u_\varepsilon)\,\frac{du_\varepsilon}{d\xi} \right), \tag{P_ε}$$
$$u_\varepsilon(\pm\infty) = u_\pm\,,$$

where $B(u)$ is a positive semidefinite $N \times N$ matrix. This approach for constructing solutions to the Riemann problem is called self-similar zero-viscosity limits. For conservative strictly hyperbolic systems, this method is known to select shocks having the internal structure of a traveling wave and to provide the unique solution to the Riemann problem for weak waves; see Tzavaras [30].

Throughout the paper we proceed under the hypothesis: There exists a family of smooth solutions u_ε to (P_ε), for $\varepsilon > 0$, that satisfy *uniform in* ε L^∞ and variation bounds,

$$|u_\varepsilon - u_-|_{L^\infty} + TV(u_\varepsilon) \le C\,, \tag{5.2}$$

as well as uniform convergence properties at infinity,

$$|u_\varepsilon(\xi) - u_\pm| \le C \exp(-\alpha\,\varepsilon) \qquad \text{for } \xi < a+1 \text{ and } \xi > b-1, \tag{5.3}$$

for some $a < b$ and $C, \alpha > 0$ independent of ε. Such solutions are constructed in LeFloch–Tzavaras [18, 19] under the following set of structural assumptions:

(i) System (1.1) is strictly hyperbolic; i.e., the matrix $A(u)$ has N real and distinct eigenvalues $\lambda_1(u) < \cdots < \lambda_N(u)$;

(ii) the initial jump $|u_+ - u_-|$ is sufficiently small;

(iii) the diffusion matrix $B(u)$ is the $N \times N$ identity matrix Id.

It is, however, expected that estimates (5.2) should hold, together with (5.3) or variants, under more general circumstances, and the analysis in this section requires only (5.2)–(5.3).

The uniform BV estimates provide a natural framework to study the notion of weak solutions for the nonconservative Riemann problem (P_0). Let $(X_\varepsilon, U_\varepsilon)$ be the arc-length reparametrization of the graph $(\xi, u_\varepsilon(\xi))$. Theorem 3.2 asserts that there exists a subsequence $\{u_{\varepsilon_n}\}$ and a generalized graph (X, U), determining a function u of bounded variation such that

$$
\begin{aligned}
&(X_{\varepsilon_n}, U_{\varepsilon_n}) \xrightarrow{\mathrm{d}} (X, U),\\
&\sigma_{\varepsilon_n}(\xi) \to \sigma(\xi)\,, \quad u_{\varepsilon_n}(\xi) \to u(\xi) \quad \text{for } \xi \in \mathcal{C}_\sigma.
\end{aligned} \tag{5.4}
$$

Recall that $\sigma_{\varepsilon_n} = X_{\varepsilon_n}^{-1}$ is a strictly increasing function, while $\sigma = X^{-1}$ is a strictly increasing multivalued map.

Using the results of sections 3 and 4, we can give a meaning to the nonconservative product $[A(u)\frac{du}{d\xi}]_{(X,U)}$, relative to the generalized graph (X, U), as a weak-$\star$ limit of $A(u_\varepsilon)\frac{du_\varepsilon}{dx}$. To this end, we use either Definition 4.1 to interpret $[A(u)\frac{du}{d\xi}]_{(X,U)}$ as a Radon measure or Definition 4.4 to define it via its distribution function F. It leads to a notion of solutions for $(\mathcal{P}_0)$ as in Definition 5.1.

DEFINITION 5.1. *Let $(X, U) : [0,1] \to [a,b]\times\mathbb{R}^N$ be a generalized graph associated with a function of bounded variation $u : [a,b] \to \mathbb{R}^N$. We say that (X, U) is a* weak solution *to the system*

$$
-\xi\,\frac{du}{d\xi} + A(u)\,\frac{du}{d\xi} = 0 \tag{5.5}
$$

if

$$
-\xi\,\frac{du}{d\xi} + \Big[A(u)\frac{du}{d\xi}\Big]_{(X,U)} = 0 \tag{5.6}
$$

in the sense of measures. Equivalently, if for any ζ, $\xi \in [a,b]$,

$$
-\Big[\xi\,u(\xi+) - \zeta\,u(\zeta-)\Big] + \int_\zeta^\xi u(\theta)d\theta + \int_{X^{-1}(\zeta-)}^{X^{-1}(\xi+)} A(U(s))\frac{dU}{ds}\,ds = 0. \tag{5.7}
$$

Remark 5.2. Relation (5.7) suggests that at points $\xi \in \mathcal{S}_{X^{-1}}$, the set where the inverse map X^{-1} is multivalued, the following analogue of the Rankine–Hugoniot conditions is satisfied:

$$
-\xi\Big[u(\xi+) - u(\xi-)\Big] + \int_{X^{-1}(\xi-)}^{X^{-1}(\xi+)} A(U(s))\frac{dU}{ds}\,ds = 0\,. \tag{5.8}
$$

Points $\xi \in \mathcal{S}_{X^{-1}}$ may correspond either to jumps or to loops.

The notion of weak solution depends on the equivalence class, but not on the specific representative, of the generalized graph.

PROPOSITION 5.3. *If a generalized graph (X, U) is a weak solution to* (5.4)*, then any path (Y, V) belonging to the same equivalence class as (X, U) is also a weak solution.*

Suppose there exists $f : \mathbb{R}^N \to \mathbb{R}^N$ such that $A = Df$. Let u of bounded variation be a solution to (5.4) *in the sense of distributions*

$$
\int_{\mathbb{R}} u\,\theta\,d\xi + \int_{\mathbb{R}} \big(\xi\,u + f(u)\,\big)\,\frac{d\theta}{d\xi}\,d\xi = 0 \tag{5.9}
$$

for every smooth function θ of compact support. Then any generalized graph associated with u is a weak solution in the sense of Definition 5.1.

The proof of Proposition 5.3 follows from the facts that the nonconservative product is independent of reparametrizations of the generalized graph (X, U), and, when $A(u) = Df(u)$, one has

$$\Big[A(u)\frac{du}{d\xi}\Big]_{(X,U)} = \frac{d}{d\xi} f(u)$$

as measures.

THEOREM 5.4. *Fix $u_\pm \in \mathbb{R}^N$. Let $u_\varepsilon : (-\infty, +\infty) \to \mathbb{R}^N$ be a family of smooth solutions to (P_ε) for $\varepsilon > 0$ that are of uniformly bounded variation and satisfy* (5.2)–(5.3). *Consider the arc-length parametrizations $(X_\varepsilon, U_\varepsilon)$ of the graphs of u_ε. There exists a subsequence $\{u_{\varepsilon_n}\}$, with $\varepsilon_n \to 0$, a generalized graph (X, U), and an associated BV function u, such that $(X_{\varepsilon_n}, U_{\varepsilon_n})$ converges to (X, U) as in* (5.4), *(X, U) is a weak solution of* (5.5), *and*

$$u(\xi) = \begin{cases} u_- & \text{for } -\infty < \xi < a + 1, \\ u_+ & \text{for } b - 1 < \xi < +\infty. \end{cases} \tag{5.10}$$

Combined with [18, 19], where the uniform bounds are established for strictly hyperbolic systems and small initial jumps $|u_+ - u_-|$, Theorem 5.4 provides an existence result for the Riemann problem (P_0). We refer to [19] for the structure of the resulting wave-fan solution of the Riemann problem and the admissibility restrictions that the process (P_ε) imposes on shocks. The relation with the solution of the Riemann problem for genuinely nonlinear systems, obtained in [14, 10], is also investigated in [19].

Proof. In view of the uniform estimates (5.2)–(5.3) and Theorem 3.2, the graphs $(X_\varepsilon, U_\varepsilon)$ converge along subsequences in the graph distance. Denote by (X, U) the limiting graph.

Observe that the right-hand side of the equation in (P_ε) tends to zero in the sense of distributions

$$\left| \varepsilon \int B(u_\varepsilon) \frac{du_\varepsilon}{d\xi} \frac{d\theta}{d\xi} \, dx \right| \le \varepsilon \, C \|\theta\|_{\mathcal{C}^1} \, TV(u_\varepsilon) \quad \longrightarrow 0 \tag{5.11}$$

for every test function θ.

To determine the limit of the right-hand side of the equation in (P_ε), one writes

$$A(u_\varepsilon) \frac{du_\varepsilon}{d\xi} = \Big[A(u_\varepsilon)\frac{du_\varepsilon}{d\xi}\Big]_{(X_\varepsilon, U_\varepsilon)}$$

and one uses the weak stability theorems, either Theorem 4.3 if the nonconservative product is viewed as a Radon measure or Theorem 4.6 if the distribution function is used instead. It follows that

$$\Big[A(u_\varepsilon)\frac{du_\varepsilon}{d\xi}\Big]_{(X_\varepsilon, U_\varepsilon)} \rightharpoonup \Big[A(u)\frac{du}{d\xi}\Big]_{(X,U)} \quad \text{weak-}\star \text{ in } \mathcal{M}[a, b].$$

Using (5.11), we conclude that (X, U) is a weak solution in the sense of Definition 5.1. Finally, the fact that $u(\xi)$ admits the boundary conditions as in (5.10) is a direct consequence of (5.3). ∎

Acknowledgment. The support of this joint project by the TMR project HCL ERBFMRXCT 960033 is gratefully acknowledged.

REFERENCES

[1] P. Aviles and Y. Giga, *Variational integrals on mappings of bounded variation and their semicontinuity*, Arch. Rational Mech. Anal., 115 (1991), pp. 201–255.

[2] L Boccardo and F. Murat, *Remarques sur l'homogénéisation de certains problèmes quasi-linéaires*, Portugal. Math., 41 (1982), pp. 535–562.

[3] A. Bressan and F. Rampazzo, *On differential systems with vector-valued impulsive controls*, Boll. Un. Mat. Ital., 7 (1988), pp. 641–656.

[4] F. Bouchut and F. James, *One-dimensional transport equations with discontinuous coefficients*, Nonlinear Anal., 32 (1997), pp. 891–933.

[5] F. Cardin and G. Zanzotto, *On continuum theories involving quasilinear first order nonconservative systems with involutions and a supplementary inequality*, Contin. Mech. Thermodyn., 3 (1991), pp. 53–63.

[6] J.F. Colombeau, *Elementary Introduction to New Generalized Functions*, North–Holland, Amsterdam, 1985.

[7] J.F. Colombeau and A.Y. Leroux, *Multiplications of distributions in elasticity and hydrodynamics*, J. Math. Phys., 29 (1988), pp. 315–319.

[8] C. Cordier, P. Degond, P. Markowich, and C. Schmeiser, *Travelling wave analysis and jump relations for Euler-Poisson model in the quasineutral limit*, Asymptot. Anal., 11 (1995), pp. 209–240.

[9] C.M. Dafermos, *Solution of the Riemann problem for a class of hyperbolic conservation laws by the viscosity method*, Arch. Rational Mech. Anal., 52 (1973), pp. 1–9.

[10] G. Dal Maso, P.G. LeFloch, and F. Murat, *Definition and weak stability of nonconservative products*, J. Math. Pures Appl., 74 (1995), pp. 483–548.

[11] G.B. Folland, *Real Analysis. Modern Techniques and Their Applications*, Wiley-Interscience, New York, 1984.

[12] T.Y. Hou and P.G. LeFloch, *Why nonconservative schemes converge to wrong solutions: error analysis*, Math. Comp., 62 (1994), pp. 497–530.

[13] P.D. Lax, *Hyperbolic systems of conservation laws* II, Comm. Pure Appl. Math., 10 (1957), pp. 537–566.

[14] P.G. LeFloch, *Entropy weak solutions to nonlinear hyperbolic systems in nonconservative form*, Comm. Partial Differential Equations, 13 (1988), pp. 669–727.

[15] P.G. LeFloch, *Shock Waves for Nonlinear Hyperbolic Systems in Nonconservative Form*, Report 593, Institute for Mathematics and Its Applications, Minneapolis, MN, 1989.

[16] P.G. LeFloch, *An existence and uniqueness result for two nonstrictly hyperbolic systems*, in Nonlinear Evolution Equations that Change Type, B. Keyfitz and M. Shearer, eds., IMA Vol. Math. Appl. 27, Springer, New York, 1990, pp. 126–138.

[17] P.G. LeFloch and T.P. Liu, *Existence theory for nonlinear hyperbolic systems in nonconservative form*, Forum Math., 5 (1993), pp. 261–280.

[18] P.G. LeFloch and A.E. Tzavaras, *Existence theory for the Riemann problem for nonconservative hyperbolic systems*, C. R. Acad. Sci., Paris Sér. I Math., 323 (1996), pp. 347–352.

[19] P.G. LeFloch and A.E. Tzavaras, in preparation.

[20] P.G. LeFloch and Z.P. Xin, *Uniqueness via the adjoint problems for systems of conservation laws*, Comm. Pure Appl. Math., 46 (1993), pp. 1499–1533.

[21] M. Marcus and V.J. Mizel, *Absolute continuity on tracks and mapping of Sobolev spaces*, Arch. Rat. Mech. Anal., 45 (1972), pp. 297–320.

[22] F. Poupaud and M. Rascle, *Measure solutions to the linear multidimensional transport equation with discontinuous coefficients*, Comm. Partial Differential Equations, 22 (1997), pp. 337–358.

[23] V.H. Ransom and D.L. Hicks, *Hyperbolic two-pressure models for two-phase flow*, J. Comput. Phys., 53 (1984), pp. 124–151.

[24] P.-A. Raviart, *Short Course on Nonlinear Hyperbolic Systems and Numerical Analysis*, Anogia Workshop, Crete, Greece, 1995, unpublished notes.

[25] P.-A. Raviart and L. Sainsaulieu, *A nonconservative hyperbolic system modeling spray dynamics,* I. *Solution of the Riemann problem*, Math. Models Methods Appl. Sci., 5 (1995), pp. 297–333.

[26] J.P. Raymond and D. Seghir, *Lower semicontinuity and integral representation of functionals*

in $BV([a,b];R^m)$, J. Math. Anal. Appl., 188 (1994), pp. 956–984.

[27] J.P. Raymond, *Definition and Stability of Nonconservative Products*, 1995, preprint.

[28] L. Sainsaulieu, *An Euler system modeling vaporizing sprays*, in Dynamics of Heterogeneous Combustion and Reacting Systems, Progress Astr. Aero., 152 (1993), pp. 280–305.

[29] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics, Heriot Watt Symposium, Vol. IV, R.J. Knops, ed., Pitman Research Notes in Math., New York, 1979, pp. 136–192.

[30] A.E. Tzavaras, *Wave interactions and variation estimates for self-similar zero-viscosity limits in systems of conservation laws*, Arch. Rational Mech. Anal., 135 (1996), pp. 1–60.

[31] A.I. Volpert, *The space BV and quasilinear equations*, Math. Sbornik, 73 (1967), pp. 225–267.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1343–1354

ROBUSTNESS OF INSTABILITY FOR THE TWO-DIMENSIONAL EULER EQUATIONS*

SUSAN FRIEDLANDER†, WALTER STRAUSS‡, AND MISHA VISHIK§

Abstract. We prove that in two dimensions any steady, inviscid, incompressible flow that is sufficiently close to an unstable flow is also unstable.

Key words. instability, inviscid fluids, Euler equations, discrete spectrum

AMS subject classifications. 76, 35

PII. S0036141098339277

Introduction. The problem of stability or instability of a steady fluid flow has been studied by many authors (see, for example, Rayleigh [1], Chandrasekhar [2], Yudovich [3], Lin [4], and Drazin and Reid [5]). In this paper we study the following question. Consider a steady flow of an inviscid, incompressible fluid, i.e., an equilibrium for the Euler equations. Assume this steady flow is (linearly) unstable. Are all Euler equilibria that are in some sense "close" to this unstable flow also unstable?

This question requires us to study the spectrum of the linearized Euler operator, which is a degenerate operator. In general the spectrum is the union of an essential spectrum and some discrete eigenvalues. The concept of a fluid Lyapunov exponent introduced by Vishik and Friedlander [6] produces an effective sufficient condition for instability in the continuous spectrum. For example, it can be used to prove that any Euler flow with a hyperbolic stagnation point is unstable (see Friedlander and Vishik [7]). There are also some equilibria where instability arises in the discrete spectrum, e.g., plane parallel shear flow with an oscillatory profile (see Friedlander, Strauss, and Vishik [8]). On the other hand, not all Euler equilibria are unstable. The celebrated Rayleigh criterion tells us that a plane parallel shear flow with no inflection points in the profile is linearly stable. This result holds for both two- and three-dimensional disturbances. The techniques of Arnold [9] prove that in two dimensions such a flow is also, in some sense, nonlinearly stable.

The stability results cited in the preceding paragraph are proved in the energy (L^2) norm. The concept of stability and instability depends very strongly on the norm. For example, Yudovich [10] proves that, in a norm that measures the derivatives of the vorticity, even a shear flow with no inflection points is unstable.

In section 1 we recall a theorem of Vishik [11], which states that the fluid Lyapunov exponent μ determines the essential spectral radius for the linearized Euler equation. Thus the flows for which μ is positive have a nonempty essential spectrum and are linearly unstable. This theorem holds in any spatial dimension for periodic boundary

*Received by the editors May 29, 1998; accepted for publication (in revised form) January 7, 1999; published electronically October 8, 1999.
http://www.siam.org/journals/sima/30-6/33927.html

†Department of Mathematics, University of Illinois at Chicago (M/C 249), 851 South Morgan Street, Chicago, IL 60607. The research of this author was partially supported by NSF grants DMS 9622563 and DMS 9300752.

‡Department of Mathematics, Brown University, Providence, RI 02912 (wstrauss@math.brown.edu). The research of this author was partially supported by NSF grant DMS 9703695.

§Department of Mathematics, University of Texas, Austin, TX 78712-1082. The research of this author was partially supported by NFS grants DMS 9531769 and DMS 9300752, and by TARP 1997-003658-071.

conditions and refers to the spectrum in L^2. Except for an occasional remark, we restrict ourselves in this paper to the case of periodic boundary conditions.

In sections 2 and 3 we prove the "robustness of instability" for ideal fluid flows in two dimensions. In two dimensions the fluid Lyapunov exponent μ is the same as the classical Lyapunov exponent. Furthermore, if a flow does not have a stagnation point, then $\mu = 0$ (see Friedlander, Strauss, and Vishik [8]). In section 2 we consider the situation where μ is positive. We prove that any two-dimensional nondegenerate flow for which $\mu > 0$ must contain a *hyperbolic* stagnation point. Hence for a nondegenerate Euler equilibrium in two dimensions, the existence of a hyperbolic stagnation point is both a necessary and sufficient condition for instability in the essential spectrum. It follows that such an instability is preserved under small perturbations of the equilibrium flow.

In section 3 we consider a two-dimensional flow that is unstable even though μ vanishes, that is, the spectrum contains some unstable eigenvalues but no unstable essential spectrum. Thus, for a given Euler equilibrium we assume the existence of an eigenvalue σ_0 with $\mathrm{Re}\sigma_0 > 0$. We prove that for any sufficiently nearby equilibrium there exists an eigenvalue σ near σ_0. To accomplish this, we first convert the eigenvalue problem for the velocity equation in L^2 to an eigenvalue problem for the vorticity equation in H^{-1}. In two dimensions the eigenvalue equation can then be converted to a condition of invertibility of an operator $(\mathrm{id} - T)$, where T is compact and depends analytically on the spectral parameter and continuously on the equilibrium. The poles of such a family of operators depend continuously on the equilibrium. This technique was developed by Guo and Strauss [12] to study the stability of BGK equilibria in collisionless plasmas. Because of the relatively simple nature of the two-dimensional vorticity equation, the proof works in two dimensions but cannot readily be extended to three.

Like most questions in fluid dynamics, the problem of the robustness of instability is much more difficult in three dimensions and in fact there is some evidence to suggest that the robustness might be false. In section 4 we exhibit a forced flow where the presence of a hyperbolic stagnation point gives a positive lower bound for the exponent μ. However, an infinitesimally small three-dimensional modification of this flow may destroy the stretching at the hyperbolic point so that the lower bound on μ collapses to zero. The behavior of the fluid exponent under perturbation is related to the problem of lower semicontinuity of the Lyapunov exponents of general dynamical systems.

In an appendix we prove the related but independent result that, for all flows with $\mu = 0$, the eigenfunctions associated with unstable eigenvalues must be C^∞.

1. The fluid Lyapunov exponent. The Euler equation for inviscid incompressible flow is

$$u_t + (u, \nabla)u + \nabla p = 0, \quad (\nabla, u) = 0.$$

The concept of a Lyapunov exponent for fluid flow was developed and utilized in a series of papers [6], [7], [10] and is summarized here. Let U be a C^∞-vector field that satisfies the steady Euler equations in a 2π-periodic domain $T^n = \mathbb{R}^n/2\pi\mathbb{Z}^n$:

$$-(U, \nabla)U - \nabla P = 0, \quad (\nabla, U) = 0, \tag{1.1}$$

where P is a C^∞-smooth pressure. We define the space of divergence-free vector fields

$$L^2_s = L^2(T^n)^n \cap \{v \mid (\nabla, v) = 0\}.$$

We consider the linearized Euler equation around U:

$$\dot{v} = -(U, \nabla)v - (v, \nabla)U - \nabla q \stackrel{\text{def}}{=} Lv, \tag{1.2}$$

$$(\nabla, v) = 0 \ , \quad v(x, 0) = v_0(x), \tag{1.3}$$

$$v_0(x) \in L^2_s. \tag{1.4}$$

The linear operator L is defined in (1.2) as acting on L^2_s with the domain

$$D(L) = \{v \in L^2_s \mid Lv \in L^2_s\}.$$

Let e^{Lt} denote the evolution operator for the linearized Euler equation. The *bicharacteristic-amplitude equations* are the system of ODEs

$$\dot{x} = U(x) \ , \quad x(0) = x_0, \tag{1.5.a}$$

$$\dot{\xi} = -\left(\frac{\partial U}{\partial x}\right)^T \xi \ , \quad \xi(0) = \xi_0, \tag{1.5.b}$$

$$\dot{b} = -\left(\frac{\partial U}{\partial x}\right) b + 2\left(\left(\frac{\partial U}{\partial x}\right) b, \xi\right) \frac{\xi}{|\xi|^2} \ , \quad b(0) = b_0. \tag{1.5.c}$$

We define the *fluid Lyapunov exponent* μ as

$$\mu \stackrel{\text{def}}{=} \mu(U) \stackrel{\text{def}}{=} \lim_{t\to\infty} \frac{1}{t} \log \sup_{\substack{|\xi_0|=1,\ |b_0|=1 \\ (\xi_0, b_0)=0}} |b(x_0, \xi_0, b_0; t)|. \tag{1.6}$$

Let $r_{\text{ess}}(e^{Lt})$ denote the essential spectral radius of the evolution operator e^{Lt} in L^2_s. It is proved in [11] that

$$\mu = \frac{1}{t} \log r_{\text{ess}}(e^{Lt}), \tag{1.7}$$

where the right side is independent of t. This identity (1.7) is valid in dimensions two and three, both for periodic boundary conditions and for equilibrium flows that are constant at infinity with perturbations in $L^2(\mathbb{R}^n)$.

The quantity μ, which can be interpreted as a Lyapunov exponent for ideal fluid flow, can be effectively computed for a number of specific flows U. For example, in [7] it is proved that any flow having exponential stretching even along a single Lagrangian trajectory has a positive μ and hence, by the theorem stated above in (1.7), is linearly unstable due to instability in the essential spectrum. On the other hand, it can be shown that $\mu = 0$ in some examples (see [8]) and hence any instability would have to occur in the discrete spectrum.

2. Perturbation of the essential spectrum for $n = 2$. Let U be a C^∞ solution to the steady Euler equation (1.1) on a torus T^2. Assume all the stagnation points (that is, points x_* where $U(x_*) = 0$) are nondegenerate (that is, $\frac{\partial U}{\partial x}(x_*)$ has both of its eigenvalues nonzero).

THEOREM 2.1. *Let V be another C^∞ solution to* (1.1) *with $\|U - V\|_{C^1(T^2)} < \varepsilon$. If the exponent μ defined in* (1.6) *is positive for the flow U and if ε is sufficiently small, then μ is also positive for the flow V.*

Proof. In two dimensions we claim that the positivity of μ for such a flow is equivalent to the existence of a hyperbolic stagnation point. To prove this claim, we appeal to results of Friedlander, Strauss, and Vishik [8, Sect. 4], concerning solutions

of the system (1.5.a)–(1.5.c) for two-dimensional flows. It is shown there that $\mu = 0$ for any flow U without a stagnation point. It is also shown that

$$|b|\,|\xi| = \text{constant} \tag{2.1}$$

along any trajectory, hence the condition $\mu > 0$ implies the existence of an exponentially decaying solution $\xi(t)$ of (1.5.b). The explicit formula for $\xi(t)$ proved in [8] is

(2.2)
$$\xi(t) = \frac{c_1 U(x(t))}{|U(x(t))|^2} + \left(c_2 - c_1 \int_0^t \left. \frac{\left(\left(\frac{\partial U}{\partial x}\right) U + \left(\frac{\partial U}{\partial x}\right)^T U, U^\perp \right)}{|U|^4} \right|_{x(\tau)} d\tau \right) U^\perp(x(t)),$$

where $\frac{\partial U}{\partial x} = U_x$ denotes the Jacobian matrix and $\perp$ is defined by $\langle u_1, u_2\rangle^\perp = \langle -u_2, u_1\rangle$. This expression is valid for any trajectory of $\dot{x} = U(x)$ *which is not itself a stagnation point*.

Since $|U|$ is bounded from above, the first term in (2.2) does not tend to zero as $t \to \infty$, so that the only way to get an exponentially decaying solution $\xi(t)$ along a sequence $t_j \to \infty$ is to have $c_1 = 0$. Thus it follows from $\mu > 0$ that for such a sequence we have $|U(x(t_j))| \le C \exp\{-\varepsilon t_j\}$ for an appropriate $\varepsilon > 0$ and $C > 0$. Because of compactness, there is a convergent subsequence $x(t_j) \to x_* \in T^2$, whence $U(x_*) = 0$. Since $x(t_j) \to x_*$, one of the eigenvalues of the Jacobian matrix $\frac{\partial U}{\partial x}(x_*)$ is negative. Since $\xi(t_j) \to 0$, one of the eigenvalues of $-(\frac{\partial U}{\partial x}(x_*))^T$ is negative. Therefore $\frac{\partial U}{\partial x}(x_*)$ has one negative and one positive eigenvalue. Thus x_* is hyperbolic. Conversely, if x_* is a hyperbolic stagnation point, then $-(\frac{\partial U}{\partial x}(x_*))^T$ has one negative eigenvalue, so that there is a solution ξ of (1.5.b) with $x(t) \equiv x_*$ that is exponentially decaying and a solution b of (1.5.c) that is exponentially growing. Hence the claim is proved.

Now we apply the Grobman–Hartmann theorem (cf. [13]) to conclude that the perturbed flow V also has a hyperbolic stagnation point for $\|U - V\|_{C^1(T^2)}$ small enough. This implies the positivity of μ for V and hence any such flow is linearly unstable. □

3. Perturbations of the discrete spectrum. Let $U \in C^\infty$ be a steady solution to the Euler equation in $T^2 = \mathbb{R}^2/2\pi\mathbb{Z}^2$:

$$-(U, \nabla)U - \nabla P = 0, \quad (\nabla, U) = 0, \qquad x \in T^2. \tag{3.1}$$

In this section we assume that $\mu = 0$ but that there is an unstable eigenvalue σ_0 (with an L^2 eigenfunction) of the generator of the corresponding linearized stability problem (1.2)–(1.4). We prove that any Euler equilibrium near U also has a nearby eigenvalue.

THEOREM 3.1. *Let $U \in C^\infty$ be a solution of* (3.1) *such that all the classical Lyapunov exponents vanish for the flow given by* (1.5.a). *Let σ_0 be an eigenvalue with* $\operatorname{Re}\sigma_0 > 0$. *Let $V \in C^\infty$ be another solution of* (3.1) *such that $\|U - V\|_{C^2(T^2)} < \varepsilon$. If ε is sufficiently small, then V also has an unstable eigenvalue σ, which is near σ_0. Thus V is linearly unstable.*

COROLLARY 3.2. *If, in addition, $\mu(U) = 0$ and U has no degenerate stagnation points, then both U and V are nonlinearly unstable.*

We denote by $\langle f\rangle$ the spatial average of a function or a distribution on T^2, by $L^2_{s,0}$ the subspace of vector fields in L^2_s with average zero, by L^2_0 the space of L^2 functions

with average zero, by H^1_0 the space of H^1 functions with average zero, and by H^{-1}_0 the subspace of H^{-1} distributions with average zero.

By *curl* we mean the scalar curl of a two-dimensional vector field: $\operatorname{curl} v = -\partial_2 v_1 + \partial_1 v_2$. If ω is a scalar function on T^2, we denote by $v = \operatorname{curl}^{-1}\omega$ the unique solution of

$$\operatorname{curl} v = \omega, \quad (\nabla, v) = 0 \quad \text{in } T^2.$$

We note that curl^{-1} maps H^{-1}_0 into $L^2_{s,0}$.

LEMMA 3.3. *Let* $\operatorname{Re}\sigma > 0$. *There is a solution of the linearized Euler equation* (1.2)–(1.4) *of the form* $e^{\sigma t}v(x)$ *with* $0 \neq v \in L^2_{s,0}$ *if and only if there is a nontrivial solution* $\omega \in H^{-1}_0$ *of the equation*

$$\sigma\omega + (U, \nabla)\omega + (\operatorname{curl}^{-1}\omega, \nabla)\operatorname{curl} U = 0. \tag{3.2}$$

Proof. Given v, we have

$$\sigma v + (U, \nabla)v + (v, \nabla)U + \nabla q = 0, \quad (\nabla, v) = 0, \quad \langle v \rangle = 0.$$

Differentiating and letting $\omega = \operatorname{curl} v$, we obtain

$$\sigma\omega + (U, \nabla)\omega + (v, \nabla)\operatorname{curl} U = 0.$$

This implies (3.2). Conversely, given ω satisfying (3.2), the function $v = \operatorname{curl}^{-1}\omega$ satisfies

$$\operatorname{curl}\,[\sigma v + (U, \nabla)v + (v, \nabla)U] = 0,$$

from which it follows that $e^{\sigma t}v(x)$ satisfies (1.2)–(1.4). ∎

It is convenient to define the operator $\Lambda = -(U, \nabla)$ with the domain $D(\Lambda) = \{\omega \in L^2_0 \mid (U, \nabla)\omega \in L^2\}$, considered as an operator on L^2_0. We denote the same operator considered on the larger space H^{-1}_0 by Λ_{-1}.

LEMMA 3.4. (a) *The operator* $e^{t\Lambda}$ *is an isometry on* L^2_0 *and its generator* Λ *has a purely imaginary spectrum.*

(b) *Assume that all the classical Lyapunov exponents vanish for the flow given by* (1.5.a). *Then* $\lim_{t\to\infty} \frac{1}{t} \log \|\exp\{t\Lambda_{-1}\}\|_{\mathcal{L}(H^{-1}_0)} = 0$ *and its generator* Λ_{-1} *also has purely imaginary spectrum.*

Proof. Let $X(x, t)$ denote the volume-preserving flow on T^2 corresponding to the vector field U (see (1.5.a)):

$$\dot{X}(x, t) = U(X(x, t)); \qquad X(x, 0) = 0, \qquad x \in T^2,\ t \in \mathbb{R}. \tag{3.3}$$

The group $\exp\{t\Lambda\}$ is given by the formula

$$(\exp\{t\Lambda\}f)(x) = f(X(x, -t)), \qquad x \in T^2,\ t \in \mathbb{R}.$$

The volume-preserving character of the flow immediately implies part (a).

We shall show that the smoothness of U guarantees that the operator $e^{t\Lambda_{-1}}$ is also well defined on H^{-1}_0. By duality it is enough to justify this statement for its adjoint Λ^*_{-1} on the space H^1_0. Indeed, $X(\cdot, t)$ is volume preserving so that for any $\psi \in H^1_0$ and $\varphi \in H^{-1}_0$ we have

$$\begin{aligned}\int_{T^2} \psi(x)\overline{(\exp\{t\Lambda_{-1}\}\varphi)(x)}\,dx &= \int_{T^2} \psi(x)\overline{\varphi(X(x, -t))}\,dx \\ &= \int_{T^2} \psi(X(x, t))\overline{\varphi(x)}\,dx = \int_{T^2} (\exp\{t\Lambda^*_{-1}\}\psi)(x)\overline{\varphi(x)}\,dx.\end{aligned}$$

Now $X(X(x,-t),t)=x$ so that we have the explicit formula for the inverse matrix $\frac{\partial X}{\partial x}(x,-t)=[\frac{\partial X}{\partial y}(X(x,-t),t)]^{-1}$. Thus the reversal of the sign of U preserves the vanishing condition on the Lyapunov exponents in a volume-preserving flow. Therefore all we have to prove is the statement for H_0^1 instead of H_0^{-1}.

Now for any $\psi\in H_0^1$ and any $t>0$,

$$\begin{aligned}\|\exp\{t\Lambda_{-1}^*\}\psi\|_{H^1}^2 &\le \int_{T^2}|\psi(X(x,t))|^2+|\nabla\psi(X(x,t))|^2\cdot\Big|\frac{\partial X(x,t)}{\partial x}\Big|^2\,dx\\ &\le C\|\psi\|_{H^1}^2\Big(1+\sup_{x\in T^2}\Big|\frac{\partial X(x,t)}{\partial x}\Big|^2\Big).\end{aligned}\tag{3.4}$$

The assumption in part (b) means that $\lim_{t\to\infty}\frac{1}{t}\log|\frac{\partial X(x,t)}{\partial x}|=0$ for all x. Thus for any $\delta>0$ and any $x\in T^2$, there is a $T(x)>0$ such that

$$\Big|\frac{\partial X(x,t)}{\partial x}\Big|\le e^{\delta t/2}\quad\text{for}\quad t\ge T(x).$$

By the smoothness of U, there is a neighborhood Q_x of x such that

$$\Big|\frac{\partial X(y,T(x))}{\partial y}\Big|\le e^{\delta T(x)},\qquad y\in Q_x.\tag{3.5}$$

By compactness, there exists a finite number of points $x_i\in T^2$ $(i=1,\dots,N)$ such that

1. $\bigcup_{i=1}^N Q_{x_i}=T^2$.
2. (3.5) is satisfied for $x=x_i$.

The inequality (3.5) now implies for *any* $y\in T^2$ and $k\in\mathbb{Z}_+$ that

$$\Big|\frac{\partial X(y,T_{i_1}+\cdots+T_{i_k})}{\partial y}\Big|\le\exp\delta(T_{i_1}+\cdots+T_{i_k}),$$

where the sequence $i_1,\dots,i_k$ is defined by the following conditions:

$$y\in U_{x_{i_1}};\ X(y,T_{x_{i_1}})\in U_{x_{i_2}};\dots;X(y,T_{x_{i_1}}+\cdots+T_{x_{i_{k-1}}})\in U_{x_{i_k}}.$$

It follows that, for *arbitrary* $t\ge 0$,

$$\Big|\frac{\partial X(y,t)}{\partial y}\Big|\le C\exp\delta t,\qquad t\ge 0,\tag{3.6}$$

where C is independent of t. Indeed, there is always a sequence $T_{i_1},\dots,T_{i_k}$ such that $T_{i_1}+\cdots+T_{i_k}\le t<T_{i_1}+\cdots+T_{i_k}+\max_i T_i$ constructed as above.

By (3.4) and (3.6), there is a $T>0$ such that

$$\|\exp\{t\Lambda_{-1}^*\}\|_{\mathcal{L}(H^1)}\le\exp\delta t,\qquad t\ge T.$$

Since $\delta>0$ is arbitrary, we deduce that the spectrum of $\exp\{t\Lambda_{-1}^*\}$ is contained in the unit circle. This proves the theorem. □

For $\operatorname{Re}\sigma>0$, we may rewrite (3.2) in terms of the resolvent of the operator Λ. We define

$$M\omega\equiv M(\sigma,U)\omega\equiv-(\sigma+(U,\nabla))^{-1}\{(\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl}U\}.\tag{3.7}$$

Thus (3.2) can be rewritten as $\omega = M\omega$.

LEMMA 3.5. (a) *M is a compact operator from H_0^{-1} to H_0^{-1}.*

(b) *M depends analytically on σ for* $\operatorname{Re}\sigma > 0$.

Proof. M is the composition of the following four continuous linear operators. The operator curl^{-1} carries H_0^{-1} into $L^2_{s,0}$. The operator of multiplication by $\nabla\operatorname{curl} U$ carries $L^2_{s,0}$ into L^2_0 because $U \in C^2$. By Lemma 3.4(a), the operator $-(\sigma+\Lambda)^{-1}$ carries L^2_0 into L^2_0. Finally, the embedding L^2_0 into H_0^{-1} is compact. This proves part (a).

The dependence on σ occurs only in the third of these operators. By Lemma 3.4(a),

$$-(\sigma+\Lambda)^{-1} = \int_0^\infty e^{-(\sigma+\Lambda)t}\, dt,$$

an integral that converges in operator norm and is analytic in the half-plane $\operatorname{Re}\sigma > 0$. Therefore M is analytic there. ∎

LEMMA 3.6. *M depends continuously on U in the following sense. Let V satisfy the same conditions as U. For any constant $C > 0$, there exists a constant C_1 such that*

$$\sup_{\operatorname{Re}\sigma \ge C} \|M(\sigma,U) - M(\sigma,V)\|_{\mathcal{L}(L^2_0)} \le C_1\ \|U-V\|_{C^2}.$$

Proof. Denote $M = M(\sigma,U)$ and $N = M(\sigma,V)$. For all $\omega \in L^2_0$, we write

$$\begin{aligned} M\omega - N\omega &= -[\sigma + (U,\nabla)]^{-1}(\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl} U \ +\ [\sigma+(V,\nabla)]^{-1}(\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl} V \\ &= I + II, \end{aligned}$$

where

$$I = [\sigma+(U,\nabla)]^{-1}(U-V,\nabla)\ [\sigma+(V,\nabla)]^{-1}\ (\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl} U$$

and

$$II = -[\sigma+(V,\nabla)]^{-1}(\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl}(U-V).$$

Now

$$\begin{aligned} \|II\|_{H_0^{-1}} &\le \|II\|_{L^2_0} \\ &\le \|U-V\|_{C^2}\ \ \|[\sigma+(V,\nabla)]^{-1}\|_{\mathcal{L}(L^2_0)}\ \|\operatorname{curl}^{-1}\omega\|_{L^2_0} \\ &\le C\|U-V\|_{C^2}\ \|\operatorname{curl}^{-1}\omega\|_{L^2_{s,0}} \\ &\le C\|U-V\|_{C^2}\ \|\omega\|_{H_0^{-1}} \end{aligned}$$

by Lemma 3.4(a). In the term I, we use Lemma 3.4(b) to estimate its first factor in the norm $\mathcal{L}(H_0^{-1}, H_0^{-1})$. Its second factor $(U-V,\nabla)\varphi$ may be written as $(\nabla,(U-V)\varphi)$. Therefore we may estimate

$$\begin{aligned} \|I\|_{H_0^{-1}} &\le C\|(U-V,\nabla)\ [\sigma+(V,\nabla)]^{-1}\ (\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl} U\|_{H_0^{-1}} \\ &\le C\|U-V\|_{L^\infty}\ \|[\sigma+(V,\nabla)]^{-1}\ (\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl} U\|_{L^2_0} \\ &\le C\|U-V\|_{L^\infty}\ \|(\operatorname{curl}^{-1}\omega,\nabla)\operatorname{curl} U\|_{L^2_0} \\ &\le C\|U-V\|_{L^\infty}\ \|U\|_{C^2}\ \|\operatorname{curl}^{-1}\omega\|_{L^2_{s,0}} \\ &\le C\|U-V\|_{L^\infty}\ \|U\|_{C^2}\ \|\omega\|_{H_0^{-1}}. \end{aligned}$$

This completes the proof of Lemma 3.6. □

Now we apply the following well-known theorem of Steinberg [14]; cf. Gohberg and Krein [15].

THEOREM 3.7. *Let $T(\sigma, s)$ be a family of compact operators on a Banach space analytic in σ and jointly continuous in (σ, s) for each $(\sigma, s) \in \Sigma \times S$, where Σ is an open set in $\mathbb{C}$ and S is an interval in $\mathbb{R}$. If for each s there exists a σ such that $I - T(\sigma, s)$ is invertible, then $(I - T(\sigma, s))^{-1}$ is meromorphic in $\sigma \in \Sigma$ for each s and the poles of $(I - T(\sigma, s))^{-1}$ depend continuously on s and can appear or disappear only at the boundary of Ω or at infinity.*

A special case of this theorem, called the analytic Fredholm theorem, which may be found in Reed and Simon [16], asserts that the set of σ such that $I - T(\sigma, s)$ is not invertible is a discrete subset of $\mathbb{C}$ and that each such σ is a pole of finite multiplicity.

We are now in a position to prove the main theorem.

Proof of Theorem 3.1. By Lemma 3.3 and (3.7), a point σ in the right half-plane is an eigenvalue of the linearized Euler equation if and only if 1 is an eigenvalue of $M(\sigma, U)$. By Theorem 3.7, the set of such σ is discrete. Now define the family of operators

$$T(\sigma, s) = (1 - s)\, M(\sigma, U) + s\, M(\sigma, V) \tag{3.8}$$

for $\operatorname{Re}\sigma > 0$ and $0 \le s \le 1$. These operators are compact on H_0^{-1}, analytic in σ, and satisfy the estimate

$$\|T(\sigma, s) - T(\sigma, 0)\| = |s|\ \|M(\sigma, U) - M(\sigma, V)\| \le C\|U - V\|_{C^2} \equiv \eta.$$

By assumption, $(I - T(\sigma, 0))^{-1}$ has a pole at some point σ_0 in the right half-plane. Fix ε_0 so small that on the circle $\Gamma = \{|\sigma - \sigma_0| = \varepsilon_0\}$ the operator $(I - T(\sigma, 0))^{-1}$, which has a discrete set of poles, exists. For η sufficiently small, $(I - T(\sigma, s))^{-1}$ also exists on the same circle Γ for all $0 \le s \le 1$. By Theorem 3.7, there exists a pole σ_1 of $(I - T(\sigma, 1))^{-1}$ within the disk $\{|\sigma - \sigma_0| < \varepsilon_0\}$. Then σ_1 is an eigenvalue for the perturbed problem with the equilibrium V. This proves the linear instability of V and completes the proof. □

Proof of Corollary 3.2. Because $\mu(U) = 0$ and U has no degenerate stagnation points, it follows from the proof of Theorem 2.1 that all the stagnation points are elliptic. So for ε sufficiently small, all the stagnation points of V are also elliptic. Therefore $\mu(V) = 0$. But we have shown that V has an unstable eigenvalue. By [8], V is *nonlinearly* unstable in the space $H_s^k = L^2_{s,0} \cap H^k(T^2)$ for any $k > 2$. □

4. A three-dimensional example. The question of the robustness of instability for steady fluid flows in three dimensions is much harder than in two dimensions and we give no answer here. The following example suggests that it may in fact not be true in complete generality since there may exist unstable flows U for which there is a stable flow arbitrarily close by.

We consider the following family of incompressible flows in $\mathbb{R}^3/2\pi\mathbb{Z}^3$ depending on a parameter $\varepsilon \in \mathbb{R}$:

$$\begin{cases} U_{1\varepsilon}(x_1, x_2, x_3) = -\sin x_3 \partial_2 \Psi(x_1, x_2), \\ U_{2\varepsilon}(x_1, x_2, x_3) = \sin x_3 \partial_1 \Psi(x_1, x_2), \\ U_{3\varepsilon}(x_1, x_2, x_3) = \varepsilon. \end{cases} \tag{4.1}$$

Although $(\nabla, U_\varepsilon) = 0$, (4.1) is not an equilibrium solution of the unforced Euler equation but only the solution of a *forced* Euler equation. However the description of the fluid Lyapunov exponent given by (1.6) still applies.

Let $\Psi \in C^\infty$ $(\mathbb{R}^2/2\pi\mathbb{Z}^2)$ and assume there exists a point (x_{01}, x_{02}), where $\nabla\Psi(x_{01}, x_{02}) = 0$ and that this point is hyperbolic for the vector field $\nabla\Psi$. The trajectory of the flow (4.1) through the point $x_0 = (x_{01}, x_{02}, x_{03})$ is the straight line on the torus $\mathbb{R}^3/2\pi\mathbb{Z}^3$

$$\begin{cases} x_1 = x_{01} & (\text{mod } 2\pi), \\ x_2 = x_{02} & (\text{mod } 2\pi), \\ x_3 = x_{03} + \varepsilon t & (\text{mod } 2\pi). \end{cases} \tag{4.2}$$

In this case equation (1.5.b) can be solved explicitly. The equation becomes

$$\dot{\xi} = -\sin(x_{03} + \varepsilon t)\ A^T \xi, \tag{4.3}$$

where

$$A = \begin{pmatrix} -\partial_{21}\Psi(x_0) & -\partial_{22}\Psi(x_0) & 0 \\ \partial_{11}\Psi(x_0) & \partial_{12}\Psi(x_0) & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Consider first the situation where $\varepsilon = 0$. Choose $\xi \equiv e_3$, which is clearly a solution to (4.3), where the vector e_j denotes a unit vector in the x_j-direction. In this case, equation (1.5.c) becomes

$$\dot{b} = -\sin x_{03}\ A\, b \quad \text{with} \quad b_3 = 0. \tag{4.4}$$

Since (x_{01}, x_{02}) is a hyperbolic point for $\nabla\Psi$, the matrix A has nonzero real eigenvalues $\pm\lambda$. Hence the exponent μ given by (1.6) has a positive bound from below. Thus such a flow, when $\varepsilon = 0$, is formally unstable.

We now turn to the case where $\varepsilon \neq 0$. The solution to (4.3) is given by

$$\xi = \exp\{-(\cos(x_{03} + \varepsilon t) - \cos x_{03})\, A^T/\varepsilon\}\ \xi_0. \tag{4.5}$$

Equation (1.5.c) becomes

$$\dot{b} = \sin(x_{03} + \varepsilon t)\ \{-A + 2|\xi|^{-2}(\xi \otimes \xi)A\}\ b. \tag{4.6}$$

This equation has the form

$$\dot{b}(t) = \dot{\varphi}(t)\ B(\varphi(t))\ b(t), \tag{4.7}$$

where $b(t)$ is vector valued, $\varphi(t) = \cos(x_{03} + \varepsilon t) - \cos x_{03}$ is a periodic scalar function, and B is a continuous matrix-valued function. The solution of (4.7) may be written as $b(t) = a(\varphi(t))$, where $da/ds = b(s)\, a(s)$ with $a(0) = b(0)$. Thus $b(t)$ is periodic. It follows that the exponent

$$\lim_{t\to\infty} \frac{1}{t} \log |b(t)| = 0.$$

Thus a fluid Lyapunov exponent that is nonzero due to stretching at the hyperbolic point when $\varepsilon = 0$ collapses to zero for a neighboring flow (4.1) with $\varepsilon \neq 0$.

Appendix A.

THEOREM A.1. *Let $U(x)$ be a C^∞ vector field on a compact n-dimensional Riemannian manifold E. Assume* $\operatorname{div} U = 0$. *Let $X(x,t)$ be the corresponding flow on E, i.e.,*

$$\dot{X}(x,t) = U(X(x,t)); \qquad X(x,0) = 0, \qquad x \in E,\ t \in \mathbb{R}.$$

Assume all forward Lyapunov exponents vanish:

$$\chi_+(x,\eta) = \overline{\lim_{t\to\infty}} \frac{1}{t} \log \left| \frac{\partial X}{\partial x}(x,t)\,\eta \right| = 0 \quad \forall\ x \in E,\ \forall\ \eta \in T_x E.$$

Let $H^k(E)$ be the Sobolev space of functions with k derivatives in $L^2(E)$. Then,

$$\lim_{t\to\infty} \frac{1}{t} \log \| \exp\{-t\,(U,\nabla)\} \|_{\mathcal{L}(H^k)} = 0.$$

Remark. As was mentioned above, the vanishing of the forward Lyapunov exponents implies the vanishing of the backward Lyapunov exponents, that is,

$$\chi_-(x,\eta) = \overline{\lim_{t\to-\infty}} \frac{1}{t} \log \left| \frac{\partial X}{\partial x}(x,t)\,\eta \right| = 0 \quad \forall\ x \in E,\ \forall\ \eta \in T_x E.$$

Proof. We use induction on k. The case $k = 1$ is proved by repeating word for word the proof of Lemma 3.4 and noting that, although $\frac{\partial X}{\partial x}$ is computed in a fixed finite system of charts, the exponential growth rate does not depend on the particular choice of this atlas. As before, $h = f \circ X(-t)$ satisfies the transport PDE

$$\frac{\partial h}{\partial t}(x,t) = -(U(x),\nabla)h, \qquad h(x,0) = f(x).$$

Now assume the theorem is proved for all H^s, $s \le k-1$. We have

$$\|h\|^2_{H^k} = \sum_{|\alpha|=k} \|D^\alpha h\|^2_{L^2} + \sum_{|\alpha|<k} \|D^\alpha h\|^2_{L^2},$$

where the second sum grows slower than any exponential. Taking β with $|\beta| = k-1$ and differentiating the equation, we arrive at

$$\frac{\partial}{\partial t}\nabla\ D^\beta h = -(U(x),\nabla)\nabla D^\beta h - \left(\frac{\partial U}{\partial x}\right)^T \nabla D^\beta h + \sum_{|\gamma|<k} b_\gamma(x) D^\gamma h,$$

with coefficients $b_\gamma(\cdot) \in C^\infty(E)$. Let $|\alpha| = k$. Using the variation of constants formula (i.e., Duhamel's principle) for solutions of an *inhomogeneous* transport equation, we get

$$\|D^\alpha h(t)\|_{L^2} \le \|G(t)\|_{L^2} \|D^\alpha h(0)\|_{L^2} + C \int_0^t \|G(t-\tau)\|_{L^2} \sum_{|\gamma|<|\alpha|} \|D^\gamma h(\tau)\|_{L^2}\, d\tau,$$

where $G(t-\tau)$ stands for the Green's function of the equation

$$\frac{\partial}{\partial t} q = -(U,\nabla)q - \left(\frac{\partial U}{\partial x}\right)^T q.$$

Again applying the induction assumption, we get for arbitrary small δ and all $t \geq 0$

$$\|G(\tau)\|_{\mathcal{L}(L^2)} \leq C_\delta \exp\{\delta t\} \quad \text{and}$$
$$\|D^\gamma h(\tau)\|_{L^2} \leq C_\delta \exp\{\delta t\} \|h(0)\|_{H^{k-1}}.$$

This implies

$$\|D^\alpha h(t)\|_{L^2} \leq C_\delta \exp\{\delta t\} \|h(0)\|_{H^k}.$$

This concludes the induction step. The proof of the theorem is complete. ∎

THEOREM A.2. *Let U be a solution to the Euler equation satisfying the conditions of Theorem* A.1. *Let σ be an eigenvalue as in Lemma* 3.3 *with* $\mathrm{Re}\sigma > 0$. *Then, the corresponding eigenfunction is in $C^\infty(E)$.*

Proof. Let $\omega \in H^{-1}$ be the eigenfunction, that is,

$$\sigma\omega + (U, \nabla)\omega + (\mathrm{curl}^{-1}\omega, \nabla)\mathrm{curl}\, U = 0. \tag{A.1}$$

Since $\mathrm{curl}^{-1}\omega \in L^2$, (A.1) implies that $\sigma\omega + (U, \nabla)\omega \in L^2$. But, according to Theorem A.1, the operator $\{\sigma + (U, \nabla)\}$ is an isomorphism on $H^k(E)$ for $\mathrm{Re}\sigma > 0$ and for integers $k \geq 0$. By duality the same statement is valid for all negative integers k as well. Since $\omega \in H^{-1}$, it follows that $\omega \in L^2$. Now by (A.1),

$$\sigma\omega + (U, \nabla)\omega \in H^1(E).$$

So by Theorem A.1, $\omega \in H^1$. Continuation of this bootstrapping argument yields

$$\omega \in \bigcap_{k=0}^{\infty} H^k(E) = C^\infty(E).$$

This proves the theorem. ∎

Remark. Under the condition that $\chi_\pm(x, \eta) \equiv 0$, it follows that any eigenfunction of M with $\mathrm{Re}\sigma > 0$ that is an arbitrary *distribution* has to be in $C^\infty(E)$. More precisely, let k be a positive integer and let $\mathrm{Re}\sigma > k\lambda$, where λ is the maximal Lyapunov exponent

$$\lambda = \lim_{t\to\infty} \frac{1}{t} \log \sup_{\pm} \sup_{(x,\eta)\in T_x E} \left| \frac{\partial X}{\partial x}(x, \pm t)\, \eta \right|.$$

Then it follows that $\omega \in H^k(E)$.

Acknowledgments. We acknowledge very helpful conversations with A. Neistadt, who suggested the idea of the example in section 4. Friedlander and Vishik thank the Mathematics Institute at Oberwolfach for the kind hospitality of its "Research in Pair" program supported by the Volkswagen Stiftung.

REFERENCES

[1] LORD RAYLEIGH, *On the stability or instability of certain fluid motions*, Proc. London Math. Soc., 9 (1880), pp. 57–70.

[2] S. CHANDRASEKHAR, *Hydrodynamic and Hydromagnetic Stability*, Oxford University Press, London, UK, 1961.

[3] V.I. YUDOVICH, *The Linearization Method in Hydrodynamical Stability Theory*, Transl. Math. Monogr. 74, AMS, Providence, RI, 1989.

[4] C.C. Lin, *The Theory of Hydrodynamic Stability*, Cambridge University Press, Cambridge, UK, 1955.
[5] P.G. Drazin and W.H. Reid, *Hydrodynamic Stability*, Cambridge Monogr. Mech. Appl. Math, Cambridge University Press, Cambridge, UK, 1981.
[6] M. Vishik and S. Friedlander, *Dynamo theory methods for hydrodynamic stability*, J. Math. Pures Appl., 72 (1993), pp. 145–180.
[7] S. Friedlander and M. Vishik, *Instability criteria for steady flows of a perfect fluid*, Chaos, 2 (1992), pp. 455–460.
[8] S. Friedlander, W. Strauss, and M. Vishik, *Nonlinear instability in an ideal fluid*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 14 (1997), pp. 187–209.
[9] V.I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, Berlin, 1978.
[10] V.I. Yudovich, *On loss of smoothness for solutions to the Euler equations*, Dynam. Contin. Media (Novosibirsk), 16 (1974), pp. 71–78 (in Russian).
[11] M. Vishik, *Spectrum of small oscillations of an ideal fluid and Lyapunov exponents*, J. Math. Pures Appl., 75 (1996), pp. 531–557.
[12] Y. Guo and W. Strauss, *Instability of periodic BGK equilibria*, Comm. Pure Appl. Math., 48 (1995), pp. 861–894.
[13] M.C. Irwin, *Smooth Dynamical Systems*, Pure Appl. Math. 94, Academic Press, Boston, MA, 1980.
[14] S. Steinberg, *Meromorphic families of compact operators*, Arch. Rational Mech. Anal., 31 (1968), pp. 372–379.
[15] I.C. Gohberg and M.G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Transl. Math. Monogr. 18, AMS, Providence, RI, 1969.
[16] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Academic Press, New York, 1972.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1355–1390

$\mathcal{A}$-QUASICONVEXITY, LOWER SEMICONTINUITY, AND YOUNG MEASURES*

IRENE FONSECA† AND STEFAN MÜLLER‡

Abstract. The notion of $\mathcal{A}$-quasiconvexity is introduced as a necessary and sufficient condition for (sequential) lower semicontinuity of

$$(u,v) \mapsto \int_\Omega f(x,u(x),v(x))\,dx$$

whenever $f : \Omega \times \mathbb{R}^m \times \mathbb{R}^d \to [0,+\infty)$ is a normal integrand, $\Omega \subset \mathbb{R}^N$ is open, bounded, $u_n \to u$ in measure, $v_n \rightharpoonup v$ in $L^p(\Omega;\mathbb{R}^d)$ ($\overset{*}{\rightharpoonup}$ if $p = +\infty$), and $\mathcal{A}v_n \to 0$ in $W^{-1,p}(\Omega)$ ($\mathcal{A}v_n = 0$ if $p = +\infty$). Here $\mathcal{A}v = \sum_{i=1}^N A^{(i)} \frac{\partial v}{\partial x_i}$ is a constant rank partial differential operator, $A^{(i)} \in \mathrm{Lin}(\mathbb{R}^d;\mathbb{R}^l)$, and $f(x,u,\cdot)$ is $\mathcal{A}$-quasi-convex if

$$f(x,u,v) \le \int_Q f(x,u,v+w(y))\,dy$$

for all $v \in \mathbb{R}^d$ and all $w \in C^\infty(Q;\mathbb{R}^d)$ such that $\mathcal{A}w = 0$, $\int_Q w(x)\,dx = 0$, and w is Q-periodic, $Q := (0,1)^N$. The characterization of Young measures generated by such sequences $\{v_n\}$ is obtained for $1 \le p < +\infty$, thus recovering the well-known results for the framework $\mathcal{A} = \mathrm{curl}$, i.e., when $v_n = \nabla\varphi_n$ for some $\varphi_n \in W^{1,p}(\Omega;\mathbb{R}^m)$, $d = N \times m$. In this case $\mathcal{A}$-quasiconvexity reduces to Morrey's notion of quasiconvexity.

Key words. $\mathcal{A}$-quasiconvexity, equi-integrability, Young measure, lower semicontinuity

AMS subject classifications. 35D99, 35E99, 49J45

PII. S0036141098339885

1. Introduction.

Recently there has been extensive research on minimization and relaxation of nonconvex energies relevant to the study of equilibria for materials exhibiting interesting, and technologically powerful, elastic and magnetic behaviors. Often a starting point for this study directly addresses minimization of the energy, leading to the search for necessary and sufficient conditions ensuring sequential weak lower semicontinuity of integrals of the form

$$(u,v) \mapsto I(u,v) := \int_\Omega f(x,u(x),v(x))\,dx,$$

where $\Omega \subset \mathbb{R}^N$ is an open, bounded set, $(u,v) : \Omega \to \mathbb{R}^m \times \mathbb{R}^d$, and $f : \Omega \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$ is a normal integrand. On the other hand, there may be situations where we need to identify $\lim_{n\to\infty} I(u_n,v_n)$ for an oscillatory sequence $\{(u_n,v_n)\}$ which does not minimize the energy. Consequently, this will entail a full characterization of the Young

*Received by the editors June 8, 1998; accepted for publication February 3, 1999; published electronically October 13, 1999.

http://www.siam.org/journals/sima/30-6/33988.html

†Department of Mathematical Sciences, Carnegie Mellon University, Pittsburgh, PA 15213 (fonseca@andrew.cmu.edu). The research of this author was partially supported by the National Science Foundation through the Center for Nonlinear Analysis, by the National Science Foundation under grant DMS–9500531, and by the Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany.

‡Max-Planck Institut für Mathematik in den Naturwissenschaften, Leipzig, Germany (sm@mis.mpg.de).

measures generated by the sequences under consideration, i.e., weak* measurable maps $\nu : \Omega \to \mathcal{P}$, where $\mathcal{P}$ is the space of probability measures on $\mathbb{R}^{m\times d}$, such that if $g : \Omega\times\mathbb{R}^m\times\mathbb{R}^d \to \mathbb{R}$ is a Carathéodory function and if $\{g(\cdot, u_n, v_n)\}$ is equi-integrable, then

$$\int_\Omega g(x, u_n(x), v_n(x))\,dx \to \int_\Omega \int_{\mathbb{R}^{m\times d}} g(x, y, z)\,d\nu_x(y, z)\,dx.$$

Although Young measures have been used for quite some time in the contexts of control theory and optimization [47], they were first introduced in a partial differential equations (PDE) framework by Tartar [41, 42, 43] in order to relate the information obtained from the linear balance equations via the method of compensated compactness with the information resulting from pointwise nonlinear constitutive relations. One application of this method was the study of quasi-linear hyperbolic equations [42], and later DiPerna [17, 18] and DiPerna and Majda [19] extended it to systems. (See [20] and [36] for further applications.) During the last few years several questions related to the study of (nonlinear) elastic materials and certain material instabilities have been successfully carried out via minimization techniques and through the understanding of the underlying Young measures [7, 8, 13, 16, 27]. Often, in this context v is the gradient ∇u of a Sobolev function $u \in W^{1,p}(\Omega;\mathbb{R}^m)$, $d = m \times N$, and coercivity of f provides boundedness of the admissible sequences in $W^{1,p}(\Omega;\mathbb{R}^m)$. If $p > 1$ then $u_n \rightharpoonup u$ in $W^{1,p}(\Omega;\mathbb{R}^m)$ (up to extraction of a subsequence). The work of Morrey [32], Ball [5], Acerbi and Fusco [1], and Marcellini [31] shows that $W^{1,p}$ (sequential) weak lower semicontinuity of

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

is equivalent to *quasiconvexity* of $f(x, u, \cdot)$ provided $0 \le f(x, u, \xi) \le a(x, u)(1 + |\xi|^p)$ for some locally bounded function $a : \Omega \times \mathbb{R}^m \to [0, +\infty)$ and for all $\xi \in \mathbb{R}^d$, almost everywhere (a.e.) $x \in \Omega$. We recall that a Borel function $f : \mathbb{M}^{m\times N} \to \mathbb{R}$ is said to be *quasi-convex* if

$$f(\xi) = \inf_{\varphi\in W_0^{1,\infty}(Q;\mathbb{R}^m)} \int_Q f(\xi + \nabla\varphi(x))\,dx, \tag{1.1}$$

where $Q := (0, 1)^N$. If f is quasi-convex then one can show that

$$f(\xi) = \inf_{\varphi\in W_{\mathrm{per}}^{1,\infty}(Q;\mathbb{R}^m)} \int_Q f(\xi + \nabla\varphi(x))\,dx, \tag{1.2}$$

where $W_{\mathrm{per}}^{1,\infty}(Q;\mathbb{R}^m)$ is the class of periodic functions in $W^{1,\infty}(Q;\mathbb{R}^m)$. Within this context, the characterization of all Young measures generated by sequences of gradients bounded in L^p was obtained by Kinderlehrer and Pedregal [28, 29]. They show that (see Theorem 2.6) in a simply connected domain Ω a weakly measurable mapping $\nu : \Omega \to \mathcal{P}$ is a Young measure generated by a sequence of gradients ∇u_n, with $\{u_n\}$ bounded in $W^{1,p}(\Omega;\mathbb{R}^m)$, if and only if three conditions are satisfied: ν is p-integrable, i.e.,

$$\int_\Omega \langle\nu_x, |\mathrm{id}|^p\rangle\,dx < +\infty;$$

the first moment $x \mapsto \langle \nu_x, \mathrm{id} \rangle$ satisfies the underlying PDE, i.e.,

$$\operatorname{curl}(\langle \nu_x, \mathrm{id} \rangle) = 0;$$

and, as suggested by (1.1), Jensen's inequality is satisfied for quasi-convex functions, i.e.,

$$\langle \nu_x, f \rangle \geq f(\langle \nu_x, \mathrm{id} \rangle)$$

for all quasi-convex functions f such that $|f(\xi)| \leq C(1 + |\xi|^p)$.

As emphasized by Tartar, in the setting of continuum mechanics and electromagnetism more general linear PDEs than $\operatorname{curl} v = 0$ arise, and the theory of compensated compactness was developed in that framework [34, 41, 42, 43, 44, 45]. To fix ideas, consider a collection of linear operators $A^{(i)} \in \mathrm{Lin}(\mathbb{R}^d, \mathbb{R}^l)$, $i = 1, \ldots, N$, and define

$$\mathcal{A}v := \sum_{i=1}^{N} A^{(i)} \frac{\partial v}{\partial x_i}, \qquad v : \mathbb{R}^N \to \mathbb{R}^d,$$

$$\mathbb{A}(w) := \sum_{i=1}^{N} A^{(i)} w_i \in \mathrm{Lin}(\mathbb{R}^d, \mathbb{R}^l), \quad w \in \mathbb{R}^N,$$

where $\mathrm{Lin(X, Y)}$ is the vector space of linear mappings from the vector space X into the vector space Y. Following Murat [34], we will assume that $\mathcal{A}$ satisfies the *constant rank* property, which states that there exists $r \in \mathbb{N}$ such that

$$\operatorname{rank} \mathbb{A}(w) = r \quad \text{for all } w \in S^{N-1}.$$

It is easy to see that the curl-free case is a particular case of this general framework (see Remark 3.3 (iii)). Other examples are discussed in Remarks 3.3 and 3.5 and in Examples 3.10 and 4.5.

We prove that a necessary and sufficient condition for weak lower semicontinuity of I, along sequences that satisfy $u_n \to u$ in measure, $v_n \rightharpoonup v$ in L^p, and $\mathcal{A}v_n \to 0$ in $W^{-1,p}(\Omega)$, is *$\mathcal{A}$-quasiconvexity* of $f(x, u, \cdot)$ (see Theorems 3.6, 3.7). The notion of $\mathcal{A}$-quasiconvexity and its implications for the lower semicontinuity of functionals $v \mapsto \int_\Omega f(v)\, dx$ were first investigated by Dacorogna, who studied, in particular, situations where the kernel of $\mathcal{A}$ contains the range of a suitable first order differential operator $\mathcal{B}$ [14, pp. 100–112] (in the general definition of $\mathcal{A}$-quasiconvexity as presented in [14, p. 13] one needs to add periodicity of the test functions to obtain necessity of $\mathcal{A}$-quasiconvexity; this leads to some difficulties in establishing sufficiency, which, under the assumption of constant rank, can be overcome using the methods presented below). Precisely and by analogy with (1.2), a function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be *$\mathcal{A}$-quasi-convex* if

$$f(v) \leq \int_Q f(v + w(x))\, dx$$

for all $v \in \mathbb{R}^d$ and all Q-periodic $w \in C^\infty(Q, \mathbb{R}^d)$ such that $\mathcal{A}(w) = 0$ and $\int_Q w(x)\, dx = 0$. In addition, we obtain the generalization to the $\mathcal{A}$-free setting of the theorem by Kinderlehrer and Pedregal concerning the characterization of gradient Young measures (see Theorem 4.1). This issue has been independently raised by Pedregal in

[35], where he studied the case of divergence-free fields (see also Remarks 3.3 (iv), 3.5 (iv)).

We remark that continuity of $\mathcal{A}$-quasiconvex functions is only guaranteed along directions in the *characteristic cone* $\Lambda := \cup_{w \in S^{N-1}} \ker \mathbb{A}(w)$, and $\mathcal{A}$-quasi-convex functions need not be (lower semi)continuous (see Proposition 3.4 and Remark 3.5 (ii)).

We note that the method used in this $\mathcal{A}$-free framework departs from the case curl-free mostly due to the lack of potential functions associated with the v_n. Indeed, in the case of gradients we reduce to the notion of quasiconvexity by localization via covering lemmas, so that on each subdomain the target function is essentially affine, followed by matching of the boundary conditions. The latter can be easily done by simple convex combinations between the potentials and the target function, avoiding layers of high concentrations of the gradients of the v_n. Clearly, the gradient of the resulting convex combinations still satisfies curl $= 0$. In the general $\mathcal{A}$-free setting, we must work directly on the v_n, and we need to find a way to project back the modified fields onto $\ker \mathcal{A}$. We perform these projections via discrete Fourier multipliers (see Lemmas 2.15, 2.16, 2.17). It is at this point that the constant rank condition enters in a crucial way. Situations where the constant rank condition fails are little understood. Tartar [41] has studied the example where $v : \mathbb{R}^2 \to \mathbb{R}^2$ and $\mathcal{A}v = (\frac{\partial v^1}{\partial x_2}, \frac{\partial v^2}{\partial x_1})$. He showed that in this case $\mathcal{A}$-quasiconvexity reduces to separate convexity, the Young measures generated by sequences along which $\{\mathcal{A}v_n\}$ is bounded in L^∞ are tensor products, and this class is strictly smaller than the class defined by duality with separately convex functions (see condition (iii) in Theorem 4.1). The class of Young measures generated by sequences that satisfy $\mathcal{A}v_n \to 0$ in $W^{-1,p}$ is not known in general (see [10, 46]). Very recently there has been progress in the case of a nonconstant rank operator $\mathcal{A}$, and this question has been completely solved by Müller for $p = 2$ (see [33]).

2. Preliminaries. In this section we recall the notion of Young measures generated by sequences bounded in L^p and by curl-free sequences. We discuss some properties of a constant rank linear partial differential operator $\mathcal{A}$, and we conclude with the decomposition lemmas, Lemmas 2.15, 2.16, and 2.17, where we show that if $\{u_n\}$ is weakly convergent in L^p and if $\mathcal{A}u_n \to 0$ in the appropriate sense then $u_n = v_n + w_n$ where $\{v_n\} \in L^p \cap \ker \mathcal{A}$ is p-equi-integrable and $\{w_n\}$ converges to zero in measure.

In the sequel $\Omega \subset \mathbb{R}^N$ is an open, bounded domain, $B(x, \varepsilon)$ denotes the open ball centered at $x \in \mathbb{R}^N$ with radius $\varepsilon > 0$, $Q := (0,1)^N$, $Q(x_0, r) := x_0 + rQ^*$, $Q^* := Q - (1/2, \dots, 1/2)$, and $S^{N-1} := \{x \in \mathbb{R}^N : |x| = 1\}$ is the unit sphere in $\mathbb{R}^N$. The Lebesgue measure in $\mathbb{R}^N$ is designated by $\mathcal{L}^N$, and H^{N-1} will stand for the $(N-1)$-dimensional Hausdorff measure in $\mathbb{R}^N$. If $1 < p < +\infty$, then $W^{-1,p}(\Omega)$ is the dual of $W_0^{1,p'}(\Omega)$, with $1/p + 1/p' = 1$, and it is well known that $F \in W^{-1,p}(\Omega)$ if and only if $F = f + \sum_{i=1}^N \frac{\partial g_i}{\partial x_i}$ in the sense of distributions for some $f, g_1, \dots, g_N \in L^p(\Omega)$. We denote by $C_0(\Omega; \mathbb{R}^d)$ the set of $\mathbb{R}^d$-valued continuous functions with compact support in Ω, endowed with the supremum norm. It is well known that the dual of the closure of $C_0(\Omega; \mathbb{R}^d)$ may be identified with the set of $\mathbb{R}^d$-valued Radon measures with finite mass, $\mathcal{M}(\Omega; \mathbb{R}^d)$, through the duality

$$\langle \mu, \varphi \rangle := \int_\Omega \varphi \cdot d\mu, \qquad \varphi \in C_0(\Omega), \mu \in \mathcal{M}(\Omega).$$

In order to simplify the notation, and when there is no ambiguity, we will abbreviate

$C_0(\Omega;\mathbb{R}^d)$ and $\mathcal{M}(\Omega;\mathbb{R}^d)$ as $C_0(\Omega)$ and $\mathcal{M}(\Omega)$, respectively. If $\mu \in \mathcal{M}(\Omega)$ and $E \subset \Omega$ is a Borel set, then $\mu\lfloor E$ stands for the restriction of the measure μ to E, i.e.,

$$\mu\lfloor E(X) := \mu(E \cap X) \quad \text{for all Borel set } X \subset \Omega.$$

We recall that given $\lambda, \mu \in \mathcal{M}(\Omega)$ with $\mu \geq 0$, by the Radon–Nikodým theorem we may decompose λ relative to μ, precisely $\lambda = \lambda_a + \lambda_s$, where λ_s and μ are *mutually singular* ($\lambda_s \perp \mu$), i.e.,

$$\lambda_s(X) = \lambda_s(X \cap B),\ \mu(X) = \mu(X \setminus B)$$

for all Borel sets $X \subset \Omega$ and for some Borel set $B \subset \Omega$, and where λ_a is *absolutely continuous with respect to* μ, $\lambda_a << \mu$, i.e., $\lambda_a(X) = 0$ whenever $X \subset \Omega$ is a Borel set and $\mu(X) = 0$. By Besicovitch's differentiation theorem we have

$$\lambda_a(X) = \int_X \frac{\partial\lambda}{\partial\mu}(x)\, d\mu, \quad \frac{\partial\lambda}{\partial\mu}(x) := \lim_{\varepsilon\to 0} \frac{\lambda(B(x,\varepsilon))}{\mu(B(x,\varepsilon))} \quad \text{for } \mu \text{ a.e. } x \in \Omega$$

and for all Borel sets $X \subset \Omega$.

If $\{z_n\}$ is a sequence bounded in $L^1(\Omega)$, then it admits a subsequence converging weakly* in the sense of measures to a measure $\mu \in \mathcal{M}(\Omega)$,

$$\int_\Omega z_{n_k}\varphi\, dx \to \int_\Omega \varphi\, d\mu$$

for all $\varphi \in C_0(\Omega)$. The *equi-integrability condition*

$$\text{for all } \varepsilon > 0 \text{ there exists } \delta > 0 \text{ such that } \mathcal{L}^N(E) < \delta \Rightarrow \sup_n \int_E |z_n(x)|\, dx < \varepsilon$$

is a necessary and sufficient condition for weak compactness in L^1 of the sequence $\{z_n\}$ (recall that Ω is bounded). If equi-integrability holds then $\mu << \mathcal{L}^N$. We will say that $\{z_n\}$ is *p-equi-integrable* if $\{|z_n|^p\}$ is equi-integrable. The following Dunford–Pettits criteria for equi-integrability are well known.

PROPOSITION 2.1. *Let $\{z_n\}$ be a sequence bounded in $L^1(\Omega)$.*

(i) *The sequence $\{z_n\}$ is equi-integrable if and only if for all $\varepsilon > 0$ there exists $M > 0$ such that*

$$\sup_n \int_{\{x\in\Omega:|z_n(x)|>M\}} |z_n(y)|\, dy < \varepsilon.$$

(ii) *The sequence $\{z_n\}$ is equi-integrable if there exists a continuous function $g : [0,+\infty) \to [0,+\infty)$ such that*

$$\lim_{t\to+\infty} \frac{g(t)}{t} = +\infty, \quad \sup_n \int_\Omega g(|z_n(x)|)\, dx < +\infty.$$

(iii) *If $\{z_n\}$ is bounded in $L^p(\Omega)$ for some $1 \leq p < +\infty$, then $\{f(z_n)\}$ is equi-integrable whenever $f : \mathbb{R}^d \to [0,+\infty)$ is a continuous function such that*

$$\lim_{|y|\to+\infty} \frac{f(y)}{|y|^p} = 0.$$

A map $\mu : E \to \mathcal{M}(\Omega)$ is said to be *weak* measurable* if $x \mapsto \langle\mu(x), \varphi\rangle$ are measurable for all $\varphi \in C_0(\Omega)$. In order to simplify the notation we denote $\mu(x)$ by μ_x.

Often the study of the behavior of solutions of nonconvex problems leads to the need to determine the limiting energy

$$\lim_{n\to\infty}\int_E f(z_n)\,dx,$$

where E is a measurable subset of Ω, $f:\mathbb{R}^d\to\mathbb{R}$ is a nonlinear function, and $\{z_n\}$ is an oscillatory sequence of measurable functions $z_n: E\to\mathbb{R}^d$. In general, the presence of oscillations entails the inequality

$$\lim_{n\to\infty}\int_E f(z_n)\,dx \neq \int_E f(z)\,dx.$$

As it turns out, the Young measure generated by (a subsequence of) $\{z_n\}$ will provide the limiting energy.

We recall that a function $f:\Omega\times\mathbb{R}^d\to\mathbb{R}$ is said to be a *normal integrand* if f is Borel measurable and $v\mapsto f(x,v)$ is lower semicontinuous for all $x\in\Omega$. Also, f is *Carathéodory* if f and $-f$ are normal integrands.

THEOREM 2.2 (fundamental theorem on Young measures [6, 11, 41]). *Let $E\subset\mathbb{R}^N$ be a measurable set of finite measure and let $\{z_n\}$ be a sequence of measurable functions, $z_n: E\to\mathbb{R}^d$. Then there exists a subsequence $\{z_{n_k}\}$ and a weak* measurable map $\nu: E\to\mathcal{M}(\mathbb{R}^d)$ such that the following hold:*

(i) *$\nu_x\geq 0$, $||\nu_x||_{\mathcal{M}(\mathbb{R}^d)}=\int_{\mathbb{R}^d}d\nu_x\leq 1$ for a.e. $x\in E$;*

(ii) *one has* (i′) *$||\nu_x||_{\mathcal{M}}=1$ for a.e. $x\in E$ if and only if*

$$\lim_{M\to\infty}\sup_k \mathcal{L}^N(\{|z_{n_k}|\geq M\})=0; \tag{2.1}$$

(iii) *if $K\subset\mathbb{R}^d$ is a compact subset and* dist $(z_{n_k},K)\to 0$ *in measure, then*

$$\operatorname{supp}\nu_x\subset K \text{ for a.e. } x\in E;$$

(iv) *if* (i′) *holds, then in* (iii) *one may replace "if" with "if and only if";*

(v) *if $f:\Omega\times\mathbb{R}^d\to\mathbb{R}$ is a normal integrand, bounded from below, then*

$$\liminf_{n\to\infty}\int_\Omega f(x,z_{n_k}(x))\,dx\geq\int_\Omega \overline{f}(x)\,dx,$$

where

$$\overline{f}(x):=\langle\nu_x,f(x,\cdot)\rangle=\int_{\mathbb{R}^d}f(x,y)\,d\nu_x(y);$$

(vi) *if* (i′) *holds and if $f:\Omega\times\mathbb{R}^d\to\mathbb{R}$ is Carathéodory and bounded from below, then*

$$\lim_{n\to\infty}\int_\Omega f(x,z_{n_k}(x))\,dx=\int_\Omega\overline{f}(x)\,dx<+\infty$$

if and only if $\{f(\cdot,z_{n_k}(\cdot))\}$ is equi-integrable. In this case

$$f(\cdot,z_{n_k}(\cdot))\rightharpoonup\overline{f}\ \text{in}\ L^1(\Omega).$$

The map $\nu: E\to\mathcal{M}(\mathbb{R}^d)$ is called the *Young measure generated by the sequence* $\{z_{n_k}\}$. It can be shown that every weak* measurable map $\nu: E\to\mathcal{M}(\mathbb{R}^d)$ that satisfies (i) is generated by some sequence $\{z_n\}$. The Young measure ν is said to be *homogeneous* if there is a Radon measure $\nu_0\in\mathcal{M}(\mathbb{R}^d)$ such that $\nu_x=\nu_0$ for a.e. $x\in E$.

Remark 2.3.

(i) Condition 2.1 holds if for some $p > 0$

$$\sup_{n\in\mathbb{N}} \int_E |z_n|^p \, dx < +\infty.$$

(ii) As a consequence of (vi), if $\{z_n\}$ is bounded in L^p and if f is a continuous function in $\mathbb{R}^d$ such that $|f(y)| \le C(1 + |y|^q)$ for some $C > 0$, $0 < q < p$, then $f(z_{n_k}) \rightharpoonup \overline{f}$ in $L^{p/q}$. Also, if $\{z_n\}$ is equi-integrable, then taking $f \equiv \mathrm{id}$ we obtain

$$z_{n_k} \rightharpoonup \overline{z} \ \text{ in } L^1(\Omega), \quad \overline{z}(x) := \langle \nu_x, \mathrm{id} \rangle.$$

PROPOSITION 2.4. *If $\{v_n\}$ generates a Young measure ν and if $w_n \to w$ in measure, then $\{v_n + w_n\}$ generates the "translated" Young measure*

$$\overline{\nu}_x := \Gamma_{w(x)} \nu_x,$$

where

$$\langle \Gamma_a \mu, \varphi \rangle := \langle \mu, \varphi(\cdot + a) \rangle$$

for $a \in \mathbb{R}^d$, $\varphi \in C_0(\mathbb{R}^d)$. In particular, if $w_n \to 0$ in measure, then $\{v_n + w_n\}$ generates the Young measure ν.

PROPOSITION 2.5. *If $\{v_n\}$ generates a Young measure ν and $u_n \to u$ a.e. in Ω, then the pair $\{(u_n, v_n)\}$ generates the Young measure μ defined by*

$$\mu_x := \delta_{u(x)} \otimes \nu_x \qquad \text{a.e. } x \in \Omega.$$

A Young measure ν is called a *gradient Young measure* if it is generated by a sequence of gradients; more precisely, ν is a $W^{1,p}$ gradient Young measure if it is generated by $\{\nabla u_n\}$ and $u_n \rightharpoonup u$ in $W^{1,p}(\Omega; \mathbb{R}^m)$. A complete characterization of such Young measures has been obtained by Kinderlehrer and Pedregal [28, 29] (see also [2, 24, 30]). A key ingredient is the notion of *quasiconvexity*: a Borel function $f : \mathbb{M}^{m\times N} \to \mathbb{R}$ is said to be *quasi-convex* if

$$f(\xi) = \inf_{\varphi \in W_0^{1,\infty}(Q;\mathbb{R}^m)} \int_Q f(\xi + \nabla\varphi(x)) \, dx.$$

If f is quasi-convex then one can show that

$$f(\xi) = \inf_{\varphi \in W^{1,\infty}_{\mathrm{per}}(Q;\mathbb{R}^m)} \int_Q f(\xi + \nabla\varphi(x)) \, dx,$$

where $W^{1,\infty}_{\mathrm{per}}(Q;\mathbb{R}^m)$ is the class of periodic functions in $W^{1,\infty}(Q;\mathbb{R}^m)$. It has been established by Morrey [32] (see also [1, 3, 5, 14, 15, 22, 23]) that sequential weak lower semicontinuity in $W^{1,p}$ and quasiconvexity are essentially equivalent. More precisely, if $0 \le f(\xi) \le C(1 + |\xi|^p)$ for some $C > 0$ and all $\xi \in \mathbb{M}^{m\times N}$ (no growth condition is necessary if $p = +\infty$), then the implication

$$u_n \rightharpoonup u \text{ in } W^{1,p} (\overset{*}{\rightharpoonup} \text{ if } p = +\infty) \Rightarrow \int_\Omega f(\nabla u) \, dx \le \liminf_{n\to\infty} \int_\Omega f(\nabla u_n) \, dx$$

holds if and only if f is quasiconvex.

THEOREM 2.6. *Let $1 \le p \le +\infty$. A weak* measurable map $\nu : \Omega \to \mathcal{M}(\mathbb{M}^{m\times N})$ is a $W^{1,p}$ gradient Young measure if and only if $\nu_x \ge 0$ a.e. $x \in \Omega$ and*

(i) *there exists* $u \in W^{1,p}(\Omega;\mathbb{R}^m)$ *such that* $\langle \nu_x, \mathrm{id}\rangle = Du$ a.e. $x \in \Omega$;
(ii) $\int_\Omega \int_{\mathbb{M}^{m\times N}} |\xi|^p \, d\nu_x(\xi)dx < +\infty$ ($\mathrm{supp}\nu_x \subset K$ *a.e.* $x \in \Omega$ *for some compact* $K \subset \mathbb{M}^{m\times N}$ *if* $p = +\infty$);
(iii) $\langle \nu_x, f\rangle \geq f(\langle \nu_x, \mathrm{id}\rangle)$ *for a.e.* $x \in \Omega$ *and for all quasiconvex* $f : \mathbb{M}^{m\times N} \to \mathbb{R}$ *(with* $|f(\xi)| \leq C(1 + |\xi|^p)$ *for some* $C > 0$ *and all* $\xi \in \mathbb{M}^{m\times N}$ *if* $1 \leq p < +\infty$*).*

Consider a collection of linear operators $A^{(i)} \in \mathrm{Lin}(\mathbb{R}^{\mathrm{d}}, \mathbb{R}^{\mathrm{l}})$, $i = 1, \ldots, N$, and define

$$\mathcal{A}v := \sum_{i=1}^{N} A^{(i)} \frac{\partial v}{\partial x_i}, \qquad v : \mathbb{R}^N \to \mathbb{R}^d,$$

$$\mathbb{A}(w) := \sum_{i=1}^{N} A^{(i)} w_i \in \mathrm{Lin}(\mathbb{R}^d, \mathbb{R}^l), \quad w \in \mathbb{R}^N,$$

where $\mathrm{Lin(X, Y)}$ is the vector space of linear mappings from the vector space X into the vector space Y.

In the sequel we will assume that $\mathcal{A}$ satisfies the *constant rank* property, namely, there exists $r \in \mathbb{N}$ such that

$$\mathrm{rank}\ \mathbb{A}(w) = r \quad \text{for all } w \in S^{N-1}. \tag{2.2}$$

Fix $w \in \mathbb{R}^N$. We define

$$\mathbb{P}(w) : \mathbb{R}^d \to \mathbb{R}^d \quad \text{to be } \textit{the orthogonal projection of } \mathbb{R}^d \text{ onto } \ker \mathbb{A}(w),$$

$$\mathbb{Q}(w) : \mathbb{R}^l \to \mathbb{R}^d, \quad \mathbb{Q}(w)\mathbb{A}(w)z := z - \mathbb{P}(w)z,\ z \in \mathbb{R}^d, \quad \mathbb{Q}(w) \equiv 0 \text{ on } (\mathrm{range}\ \mathbb{A}(w))^{\perp}.$$

Proposition 2.7. *If* (2.2) *holds then the map* $\mathbb{P} : \mathbb{R}^N \setminus \{0\} \to \mathrm{Lin}(\mathbb{R}^{\mathrm{d}}; \mathbb{R}^{\mathrm{d}})$ *is smooth and homogeneous of degree zero, and the map* $\mathbb{Q} : \mathbb{R}^N \setminus \{0\} \to \mathrm{Lin}(\mathbb{R}^{\mathrm{l}}; \mathbb{R}^{\mathrm{d}})$ *is smooth and homogeneous of degree* -1.

Let $\Delta := \mathbb{Z}^N$ be the *unit lattice*, i.e., the additive group of points in $\mathbb{R}^N$ with integer coordinates. We say that $f : \mathbb{R}^N \to \mathbb{R}^d$ is Δ-*periodic* if

$$f(x + \lambda) = f(x) \quad \text{for all } x \in \mathbb{R}^N, \lambda \in \Delta.$$

A Δ-periodic function f may be identified with a function f_T on the N *torus*

$$T_N := \{(e^{2\pi i x_1}, \ldots, e^{2\pi i x_N}) \in \mathbb{C}^N : (x_1, \ldots, x_N) \in \mathbb{R}^N\}$$

through the relation

$$f_T(e^{2\pi i x_1}, \ldots, e^{2\pi i x_N}) := f(x_1, \ldots, x_N).$$

The space $L^p(T_N)$ is identified with $L^p(Q)$, and $C(T_N)$ is the set of Δ-periodic continuous functions on $\overline{Q}$.

Proposition 2.8 (see [9]). *Let* $w \in L^p(T_N; \mathbb{R}^d)$, $1 \leq p \leq +\infty$, *and set* $w_n(x) := w(nx), n \in \mathbb{N}$. *If* $E \subset \mathbb{R}^N$ *is a measurable set, then*

$$w_n \rightharpoonup \int_{T_N} w(y)\, dy \quad \textit{in } L^p(E; \mathbb{R}^d) \quad (\overset{*}{\rightharpoonup} \textit{ if } p = +\infty).$$

In particular, $\{w_n\}$ *generates the homogeneous Young measure* $\nu := \overline{\delta_w}$, *where*

$$\langle \overline{\delta_w}, \varphi \rangle := \int_{T_N} \varphi(w(y))\, dy \quad \text{for all } \varphi \in C_0(\mathbb{R}^d).$$

We recall some results on Fourier transforms of periodic functions (see [38, 39]). If $f \in L^1(T_N)$, then its *Fourier coefficients* are defined as

$$\hat{f}(\lambda) := \int_{T_N} f(x) \mathrm{e}^{-2\pi i x\cdot\lambda}\, dx, \qquad \lambda \in \Delta,$$

and the following hold.

THEOREM 2.9.

(i) *The trigonometric polynomials*

$$R(x) := \sum_{\lambda\in\Delta'} a_\lambda e^{-2\pi i x\cdot\lambda}, \quad \Delta' \text{finite subset of } \Delta, a_\lambda \in \mathbb{C},$$

are dense in $C(T_N)$ *and in* $L^p(T_N)$ *for all* $1 \le p < +\infty$.

(ii) *If* $\mu \in \mathcal{M}(T_N)$ *and* $\langle \mu, e^{-2\pi i x\cdot\lambda}\rangle = 0$ *for all* $\lambda \in \Delta$ *then* $\mu \equiv 0$.

(iii) *If* $f \in L^2(T_N)$ *then*

$$f(x) = \sum_{\lambda\in\Delta} \hat{f}(\lambda) e^{2\pi i x\cdot\lambda}, \quad \sum_{\lambda\in\Delta} |\hat{f}(\lambda)|^2 = ||f||_{L^2}.$$

COROLLARY 2.10. *If* $f \in L^1(T_N)$ *and* $\sum_{\lambda\in\Delta} |\hat{f}(\lambda)| < +\infty$, *then there exists a representative* $\overline{f}$ *of* f *such that* $\overline{f} \in C(T_N)$ *and for all* $x \in T_N$

$$\overline{f}(x) = \sum_{\lambda\in\Delta} \hat{f}(\lambda) e^{2\pi i x\cdot\lambda}.$$

COROLLARY 2.11. *If* $f \in C^k(T_N)$ *for some* $k > N/2$ *then*

$$\sum_{\lambda\in\Delta} |\hat{f}(\lambda)| < +\infty.$$

Let $(L^p(T_N), L^q(T_N))$ denote the class of (p, q) *Fourier multiplier operators*, i.e., the class of all bounded linear operators $T : L^p(T_N) \to L^q(T_N)$ which commute with translations,

$$\Gamma_h T = T \Gamma_h \quad \text{for all } h \in \mathbb{R}^N,$$

where $\Gamma_h f(x) := f(x - h)$.

THEOREM 2.12. *If* $1 \le p, q \le +\infty$ *and if* $T \in (L^p(T_N), L^q(T_N))$, *then there exists a bounded function* $\Theta : \Delta \to \mathbb{C}$ *such that*

$$Tf(x) := \sum_{\lambda\in\Delta} \Theta(\lambda)\hat{f}(\lambda) e^{2\pi i x\cdot\lambda} \quad \text{if } f \in L^p(T_N) \text{is given by } f(x) = \sum_{\lambda\in\Delta} \hat{f}(\lambda) e^{2\pi i x\cdot\lambda}.$$

The collection of coefficients $\{\Theta(\lambda)\}_{\lambda\in\Delta}$ *is called the Fourier multiplier associated with* T.

It can be shown that a certain class of continuous functions on the unit sphere S^{N-1} are Fourier multipliers. Precisely (see [38, Example iii, pp. 94], [39, Corollary 3.16, p. 263, and remark just below]),

PROPOSITION 2.13. *If Θ is homogeneous of degree zero and if it is infinitely differentiable on S^{N-1}, then the operator $T_\Theta : L^p(T_N) \to L^p(T_N)$ defined by*

$$T_\Theta f(x) := \sum_{\lambda \in \Delta \setminus \{0\}} \Theta(\lambda) \hat{f}(\lambda) e^{2\pi i x \cdot \lambda} \text{ if } f \in L^p(T_N) \text{ is given by } f(x) = \sum_{\lambda \in \Delta} \hat{f}(\lambda) e^{2\pi i x \cdot \lambda}$$

is a Fourier multiplier operator for $1 < p < +\infty$.

If (2.2) holds, then in light of Propositions 2.7 and 2.13 the functions

$$\Theta_{ij} : w \in \mathbb{R}^N \mapsto \mathbb{P}(w)_{ij}, \quad i, j \in \{1, \dots, d\},$$

generate the Fourier multipliers $\{\Theta_{ij}(\lambda)\}_{\lambda \in \Delta \setminus \{0\}}$ associated with the Fourier multiplier operators $T_{\Theta_{ij}}$, and we define the operators

$$(\mathbb{T}u)_i(x) := (T_{\Theta_{ij}} u_j)(x) \quad \text{for } u \in L^p(T_N; \mathbb{R}^d), \quad i = 1, \dots, N,$$

where the summation convention for repeated indices is used.

LEMMA 2.14. *Suppose that (2.2) holds and let $1 < p < +\infty$. Then*

(i) *$\mathbb{T} : L^p(T_N; \mathbb{R}^d) \to L^p(T_N; \mathbb{R}^d)$ is a linear, bounded operator that vanishes on constant mappings;*

(ii) *if $u \in L^p(T_N; \mathbb{R}^d)$ then $\mathbb{T} \circ \mathbb{T}u = \mathbb{T}u$, and $\mathcal{A}(\mathbb{T}u) = 0$;*

(iii) *$||u - \mathbb{T}u||_{L^p} \le C_p ||\mathcal{A}u||_{W^{-1,p}}$ for all $u \in L^p(T_N; \mathbb{R}^d)$, such that $\int_{T_N} u \, dx = 0$, and for some $C_p > 0$;*

(iv) *suppose that $\{u_n\}$ is a sequence bounded in $L^p(T_N; \mathbb{R}^d)$ and $\{|u_n|^p\}$ is equi-integrable. Then $\{|\mathbb{T}u_n|^p\}$ is still equi-integrable.*

Proof. Property (i) follows from the definition of $\mathbb{T}$ and from Propositions 2.7 and 2.13. Property (ii) is an immediate consequence of the fact that $\mathbb{P}$ is a projection.

To prove (iii), we note that by Corollaries 2.10 and 2.11 for $u \in C^\infty(T_N; \mathbb{R}^d)$, with $\int_{T_N} u \, dx = 0$, we have

$$\begin{aligned} u - \mathbb{T}u &= \Sigma_{\lambda \in \Delta \setminus \{0\}} \, \mathbb{Q}(\lambda) \mathbb{A}(\lambda) \hat{u}(\lambda) e^{2\pi i x \cdot \lambda} \\ &= \Sigma_{\lambda \in \Delta \setminus \{0\}} \, \mathbb{Q}\left(\frac{\lambda}{|\lambda|}\right) \mathbb{A}\left(\frac{\lambda}{|\lambda|}\right) \hat{u}(\lambda) e^{2\pi i x \cdot \lambda}, \end{aligned}$$

where we have used the linearity of $\mathbb{A}$ and the fact that $\mathbb{Q}$ is homogeneous of degree -1. By Proposition 2.7 the inequality in (iii) is obtained, and the result for L^p periodic functions with zero average follows via a density argument.

To prove (iv) consider the truncation $\tau_\alpha : \mathbb{R}^d \to \mathbb{R}^d$ given by

$$\tau_\alpha(z) := \begin{cases} z & \text{if} \quad |z| \le \alpha, \\ \alpha \frac{z}{|z|} & \text{if} \quad |z| > \alpha. \end{cases}$$

Since $\{\tau_\alpha u_n\}$ is bounded in L^∞ we have that$\{\mathbb{T}(\tau_\alpha u_n)\}$ is bounded in L^q for all $p \le q < +\infty$, and so $\{|T(\tau_\alpha u_n)|^p\}$ is equi-integrable. On the other hand, by the equi-integrability of $\{u_n\}$ we have that

$$\lim_{\alpha \to \infty} \sup_n ||u_n - \tau_\alpha u_n||_p = 0,$$

and by (i) we conclude that

$$\lim_{\alpha\to\infty}\sup_n ||\mathbb{T}(u_n-\tau_\alpha u_n)||_p=0,$$

and the assertion is proved. □

We note that, with the exception of Lemma 2.14 (iv), the above closely follows Murat's work (see [34]).

Decomposition results similar to the ones obtained below may be found in [24] and [30] in the particular case of curl-free fields.

LEMMA 2.15 ($1<p<+\infty$). *Let $1<p<+\infty$, let $\{u_n\}$ be a bounded sequence in $L^p(\Omega;\mathbb{R}^d)$ such that $\mathcal{A}u_n\to 0$ in $W^{-1,p}(\Omega)$, $u_n\rightharpoonup u$ in $L^p(\Omega;\mathbb{R}^d)$, and assume that $\{u_n\}$ generates the Young measure ν. Then there exists a p-equi-integrable sequence $\{v_n\}\subset L^p(\Omega;\mathbb{R}^d)\cap\ker\mathcal{A}$ such that*

$$\int_\Omega v_n\,dx=\int_\Omega u\,dx,\quad ||v_n-u_n||_{L^q(\Omega)}\to 0\quad \textit{for all } 1\le q<p$$

and, in particular, $\{v_n\}$ still generates ν.

Proof. After an affine rescaling, we may suppose that $\Omega\subset Q$. The assumptions imply that $\mathcal{A}u=0$, and by linearity (and Proposition 2.4) we may take $u=0$. By Theorem 2.2 (v) we have

$$\int_\Omega\int_{\mathbb{R}^d}|z|^p\,d\nu_x(z)\,dx<+\infty$$

and so, using Theorem 2.2 (vi), we obtain

$$\begin{aligned}\lim_{k\to\infty}\lim_{n\to\infty}\int_\Omega|\tau_k(u_n)|^p\,dx&=\lim_{k\to\infty}\int_\Omega\int_{\mathbb{R}^d}|\tau_k(z)|^p\,d\nu_x(z)\,dx\\&=\int_\Omega\int_{\mathbb{R}^d}|z|^p\,d\nu_x(z)\,dx.\end{aligned}$$

Therefore we may find an increasing sequence $\alpha_n\to+\infty$ such that the truncated sequence $\{\tau_{\alpha_n}\circ u_n\}$ satisfies

$$\lim_{n\to\infty}\int_\Omega|\tau_{\alpha_n}\circ u_n|^p\,dx=\int_\Omega\int_{\mathbb{R}^d}|z|^p\,d\nu_x(z)\,dx.\tag{2.3}$$

On the other hand, as $\{u_n\}$ is equi-integrable,

$$\tau_{\alpha_n}\circ u_n-u_n\to 0\quad\text{in measure and weakly in } L^p(\Omega).$$

Thus, by Proposition 2.4, the sequence $\{\tilde u_n\}:=\{\tau_{\alpha_n}\circ u_n\}$ still generates the Young measure ν. By Theorem 2.2 (vi) and (2.3) we conclude that $\{\tilde u_n\}$ is p-equi-integrable. Moreover, if $1<q<p$, then

$$\begin{aligned}||\tilde u_n-u_n||^q_{L^q(\Omega)}&\le\int_{\{|u_n|\ge\alpha_n\}}2^q|u_n|^q\,dx\\&\le\alpha_n^{q-p}\,2^q\int_{T_N}|u_n|^p\,dx\to 0\quad\text{as } n\to+\infty,\end{aligned}$$

and thus $\mathcal{A}\tilde{u}_n \to 0$ in $W^{-1,q}(\Omega)$. Also, by virtue of the compact imbedding $L^q(\Omega) \hookrightarrow W^{-1,q}(\Omega)$, we have for all $\varphi \in C_0^\infty(\Omega;[0,1])$

$$\mathcal{A}(\varphi\tilde{u}_n) = \varphi\mathcal{A}(\tilde{u}_n) + \sum_{i=1}^{N} A^{(i)}(\tilde{u}_n)\,\frac{\partial\varphi}{\partial x_i} \to 0 \text{ in } W^{-1,q}(\Omega).$$

Thus we may select a sequence $\{\varphi_n\} \subset C_0^\infty(\Omega;[0,1])$ with $\varphi_n \nearrow 1$, and such that, setting $\hat{u}_n := \varphi_n\,\tilde{u}_n$, $\{\hat{u}_n\}$ is p-equi-integrable,

$$\hat{u}_n \rightharpoonup 0 \quad \text{in } L^p(\Omega),\ \mathcal{A}\hat{u}_n \to 0 \text{ in } W^{-1,q}(\Omega).$$

Extend $\hat{u}_n$ by zero to $Q \setminus \Omega$ and then periodically. We define

$$\tilde{v}_n := \mathbb{T}\left(\hat{u}_n - \int_{T_N} \hat{u}_n\,dy\right).$$

By Lemma 2.14 (iv) the sequence $\{\tilde{v}_n\} \subset L^p(\Omega;\mathbb{R}^d) \cap \ker\mathcal{A}$ is p-equi-integrable, and we have

$$\begin{aligned}(2.4)\qquad ||\tilde{v}_n - u_n||_{L^q(\Omega)} &\le ||\tilde{v}_n - \tilde{u}_n||_{L^q(\Omega)} + ||\tilde{u}_n - u_n||_{L^q(\Omega)}\\ &\le ||\tilde{v}_n - \hat{u}_n||_{L^q(\Omega)} + ||\hat{u}_n - \tilde{u}_n||_{L^q(\Omega)} + ||\tilde{u}_n - u_n||_{L^q(\Omega)}\\ &=: I_1^n + I_2^n + I_3^n.\end{aligned}$$

We have already seen that $I_3^n \to 0$ as $n \to \infty$, and the p-equi-integrability of $\{\tilde{u}_n\}$ entails

$$\lim_{n\to\infty} I_2^n = 0.$$

Using Lemma 2.14 (iii) and the fact that $\int_{T_N} \hat{u}_n\,dy \to 0$, we obtain

$$\begin{aligned}\lim_{n\to\infty} I_1^n &\le \lim_{n\to\infty} \left|\left|\hat{u}_n - \int_{T_N} \hat{u}_n\,dy - \mathbb{T}\left(\hat{u}_n - \int_{T_N} \hat{u}_n\,dy\right)\right|\right|_{L^q(T_N)}\\ &\le \lim_{n\to\infty} C_q\,||\mathcal{A}\hat{u}_n||_{W^{-1,q}(T_N)}\\ &= 0.\end{aligned}$$

In particular, by Proposition 2.4 $\{\tilde{v}_n\}$ still generates ν. Finally, it suffices to set

$$v_n := \tilde{v}_n - \int_\Omega \tilde{v}_n\,dy.$$

Note that if the initial sequence $\{u_n\}$ is p-equi-integrable, then there is no need to construct the truncated sequence $\{\tilde{u}_n\}$, and from (2.4) it follows that $||v_n - u_n||_{L^p(T_N)} \to 0$. □

LEMMA 2.16 ($p = 1$). *Let $\{u_n\}$ be a sequence converging weakly in $L^1(\Omega;\mathbb{R}^d)$ to a function u, $\mathcal{A}u_n \to 0$ in $W^{-1,r}(\Omega_N)$ for some $r \in (1, N/(N-1))$, and assume that $\{u_n\}$ generates a Young measure ν. Then there exists an equi-integrable sequence $\{v_n\} \in L^1(\Omega;\mathbb{R}^d) \cap \ker\mathcal{A}$ such that*

$$\int_\Omega v_n\,dx = \int_\Omega u\,dx, \quad ||v_n - u_n||_{L^1(\Omega)} \to 0$$

and, in particular, $\{v_n\}$ still generates ν.

Proof. The proof is similar to the one given above, and once again we may assume that $\Omega \subset Q$ and $u = 0$. Due to the equi-integrability of $\{u_n\}$ we do not need to truncate the sequence, so we set $\tilde{u}_n := u_n$. Also, by mollification we may assume that $\hat{u}_n \in C_0^\infty(\Omega;\mathbb{R}^d)$, where in the diagonalization argument leading to the construction of $\hat{u}_n$ we use the compact imbedding $L^1(\Omega) \hookrightarrow W^{-1,r}(\Omega)$. We have

$$||\tilde{v}_n - u_n||_{L^1(\Omega)} \leq ||\tilde{v}_n - \hat{u}_n||_{L^1(\Omega)} + ||\hat{u}_n - u_n||_{L^1(\Omega)}$$

and the last term on the right-hand side converges to zero due to the equi-integrability of $\{u_n\}$. Finally,

$$\begin{aligned}\lim_{n\to\infty} ||\tilde{v}_n - \hat{u}_n||_{L^1(\Omega)} &\leq \lim_{n\to\infty} C_r \left|\left|\hat{u}_n - \int_{T_N} \hat{u}_n \, dy - \mathbb{T}\left(\hat{u}_n - \int_{T_N} \hat{u}_n \, dy\right)\right|\right|_{L^r(T_N)} \\ &\leq C_r \lim_{n\to\infty} ||\mathcal{A}\hat{u}_n||_{W^{-1,r}(T_N)} \\ &= 0,\end{aligned}$$

where we have used the fact that $\int_{T_N} \hat{u}_n \, dy \to 0$. ∎

LEMMA 2.17 ($p = +\infty$). *Let $\{u_n\}$ be a sequence that satisfies $u_n \overset{*}{\rightharpoonup} u$ in $L^\infty(T_N;\mathbb{R}^d)$, $\mathcal{A}u_n \rightharpoonup 0$ in $L^p(T_N)$ for some $p > N$, and assume that $\{u_n\}$ generates a Young measure ν. Then there exists a sequence $\{v_n\} \in L^\infty(T_N;\mathbb{R}^d) \cap \ker \mathcal{A}$ such that*

$$\int_{T_N} v_n \, dx = \int_{T_N} u \, dx, \quad ||v_n - u_n||_{L^\infty(T_N)} \to 0$$

and, in particular, $\{v_n\}$ still generates ν.

Proof. As before assume that $u = 0$ and set

$$v_n := \mathbb{T}\left(u_n - \int_{T_N} u_n \, dy\right).$$

Since $\int_{T_N} u_n \, dy \to 0$, we have

$$\sup_{n\in\mathbb{N}} ||v_n - u_n||_{W^{1,p}(T_N)} \leq C_p \sup_{n\in\mathbb{N}} ||\mathcal{A}u_n||_{L^p(T_N)} < +\infty$$

and

$$\lim_{n\to\infty} ||v_n - u_n||_{L^p(T_N)} \leq C_p \lim_{n\to\infty} ||\mathcal{A}u_n||_{W^{-1,p}(T_N)} = 0,$$

and we conclude that the functions $v_n - u_n$ converge to zero uniformly. ∎

The last result of this section will enable us in section 4 to focus our attention on the characterization of $\mathcal{A}$-1-Young measures, where a Young measure ν is said to be a *$\mathcal{A}$-p-Young measure* if it is generated by a sequence in $\ker \mathcal{A}$ which is weakly convergent in $L^p(\Omega)$.

COROLLARY 2.18. *Let $1 < p < +\infty$. If ν is a $\mathcal{A}$-1-Young measure with*

$$\int_\Omega \int_{\mathbb{R}^d} |z|^p \, d\nu_x(z) \, dx < +\infty,$$

then ν is a $\mathcal{A}$-p-Young measure generated by a p-equi-integrable sequence.

Proof. Assume that ν is generated by an equi-integrable sequence $\{u_n\} \subset L^1(\Omega) \cap \ker \mathcal{A}$, and

$$\int_\Omega \int_{\mathbb{R}^d} |z|^p \, d\nu_x(z) \, dx < +\infty.$$

Following the beginning of the proof of Lemma 2.15, we may find a sequence of truncations $\{\tilde{u}_n\} \subset \ker \mathcal{A}$, bounded in $L^p(\Omega; \mathbb{R}^d)$, that still generates ν since, by equi-integrability,

$$||\tilde{u}_n - u_n||_{L^1(\Omega)} \to 0.$$

The result now follows by direct application of Lemma 2.15 to the sequence $\{\tilde{u}_n\}$. □

3. $\mathcal{A}$-quasi-convexity: A necessary and sufficient condition for lower semicontinuity. Using the notation introduced in section 2, consider an operator $\mathcal{A}$ satisfying the constant rank property (2.2). In this section we will prove lower semicontinuity of functionals with normal integrands with respect to weakly convergent sequences with weak limits in the kernel of $\mathcal{A}$. In what follows Ω is a bounded, open subset of $\mathbb{R}^N$.

Given a normal integrand $f : \Omega \times \mathbb{R}^m \times \mathbb{R}^d \to \mathbb{R}$, we define

$$I(u, v) := \int_\Omega f(x, u(x), v(x)) \, dx$$

for measurable $(u, v) : \Omega \to \mathbb{R}^m \times \mathbb{R}^d$.

DEFINITION 3.1. *A function $f : \mathbb{R}^d \to \mathbb{R}$ is said to be* $\mathcal{A}$-quasi-convex *if*

$$f(v) \leq \int_Q f(v + w(x)) \, dx$$

for all $v \in \mathbb{R}^d$ and all $w \in C^\infty(T_N; \mathbb{R}^d)$ such that $\mathcal{A}(w) = 0$ and $\int_{T_N} w(x) \, dx = 0$.

DEFINITION 3.2. *Given a Borel function $f : \mathbb{R}^d \to \mathbb{R}$ we define the* $\mathcal{A}$-quasi-convex envelope *of f at $v \in \mathbb{R}^d$ as*

$$Q_{\mathcal{A}} f(v) := \inf \left\{ \int_{T_N} f(v + w(x)) \, dx : w \in C^\infty(T_N) \cap \ker \mathcal{A}, \int_{T_N} w \, dx = 0 \right\}. \tag{3.1}$$

Clearly $f = Q_{\mathcal{A}} f$ when f is *$\mathcal{A}$-quasi-convex*.

Remark 3.3.

(i) It follows immediately from Jensen's inequality that convex functions are $\mathcal{A}$-quasi-convex.

(ii) If f is upper semicontinuous and locally bounded from above, then $C^\infty(T_N)$ may be replaced by $L^\infty(T_N)$ in Definition 3.1. Indeed, it suffices to approximate a given function $w \in L^\infty(T_N) \cap \ker \mathcal{A}$, with $\int_{T_N} w \, dx = 0$, by the mollified sequence

$$w_\varepsilon := \rho_\varepsilon * w - \int_{T_N} \rho_\varepsilon * w \, dy,$$

where $w_\varepsilon \in C^\infty(T_N) \cap \ker \mathcal{A}$, are Q-periodic, and have zero average. The result now follows by Fatou's lemma. If, in addition, $|f(v)| \leq C(1 + |v|^p)$ for some $C > 0$ and all $v \in \mathbb{R}^d$, then $C^\infty(T_N)$ may be replaced by $L^p(T_N)$ in (3.1).

(iii) Given a matrix-valued function $V : \Omega \subset \mathbb{R}^N \to \mathbb{M}^{m\times n} \equiv \mathbb{R}^d$, $d := mn$, $n = N + \rho$, $\rho \geq 0$, we write

$$V = (F \mid \xi), \quad F \in \mathbb{M}^{m\times N},\ \xi \in \mathbb{M}^{m\times \rho},$$

where F is the matrix of the first N columns of V, and ξ is the matrix of the remaining ρ columns. In the context of membrane or film theories, $N = 2$, $m = 3$, $\rho = 1$, and F is the gradient of the membrane deformation. In the context of general nonlinear elasticity, $N = m = 3, \rho = 0$, and F is the gradient of the deformation of the elastic solid. The underlying PDE is then

$$\operatorname{curl} F = 0, \quad \text{i.e.,}\ \frac{\partial F_{jk}}{\partial x_i} - \frac{\partial F_{ji}}{\partial x_k} = 0, \quad 1 \leq j \leq m, 1 \leq i, k \leq N.$$

We may rewrite these PDEs as $\mathcal{A}V = 0$, where $l := N^2 m$ and

$$A^{(r)}_{(j,k,i),(q,p)} := \delta_{ri}\delta_{qj}\delta_{pk} - \delta_{rk}\delta_{qj}\delta_{pi}, \quad 1 \leq j, q \leq m,\ 1 \leq i, k, p, r \leq N,$$

$$A^{(r)}_{(j,k,i),(q,p)} = 0 \quad \text{if } p = N + 1, \ldots, n.$$

The constant rank condition (2.2) is satisfied, since $\dim(\ker \mathbb{A}(w)) = m + m \times \rho$ for all $w \in S^{N-1}$. Indeed,

$$\begin{aligned}\ker \mathbb{A}(w) &= \{V \in \mathbb{M}^{m\times n} : \mathbb{A}(w)V = 0\}\\ &= \{V = (F \mid \xi) \in \mathbb{M}^{m\times n} : w_i F_{jk} - w_k F_{ji} = 0,\ 1 \leq j \leq m, 1 \leq i, k \leq N\}\\ &= \{V = (F \mid \xi) \in \mathbb{M}^{m\times n} : F = a \otimes w \text{ for some } a \in \mathbb{R}^m\}.\end{aligned}$$

When $\rho = 0$ and f is locally bounded, then (3.1) reduces to the usual *quasi-convex envelope* of f,

$$\begin{aligned}Q_{\mathcal{A}} f(v) &:= \inf\left\{\int_{T_N} f(v + \nabla\varphi(x))\,dx : \varphi \in C^\infty(T_N; \mathbb{R}^m)\right\}\\ &= \inf\left\{\int_Q f(v + \nabla\varphi(x))\,dx : \varphi \in C_0^\infty(Q; \mathbb{R}^m)\right\}.\end{aligned}$$

(iv) Now we consider the div-free case (see also [35]). Here $d = N, l = 1$,

$$A^{(i)}_j := \delta_{ij},$$

so that

$$\mathcal{A}u = 0 \quad \text{if and only if } \operatorname{div} u = 0.$$

Once again, the constant rank condition (2.2) holds, as for all $w \in S^{N-1}$

$$\begin{aligned}\ker \mathbb{A}(w) &= \left\{v \in \mathbb{R}^N : \sum_{i=1}^N A^{(i)} w_i(v) = 0\right\}\\ &= \{v \in \mathbb{R}^N : v \cdot w = 0\}.\end{aligned}$$

Therefore $\dim(\ker \mathbb{A}(w)) = N - 1$.

PROPOSITION 3.4. *If $f : \mathbb{R}^d \to \mathbb{R}$ is upper semicontinuous, then $Q_{\mathcal{A}}f$ is $\mathcal{A}$-quasi-convex and upper semicontinuous. Moreover, the restriction of $Q_{\mathcal{A}}f$ to each cone $a + \Lambda$, $a \in \mathbb{R}^d$, is convex, i.e.,*

$$Q_{\mathcal{A}}f(\theta y + (1-\theta)z) \le \theta Q_{\mathcal{A}}f(y) + (1-\theta)Q_{\mathcal{A}}f(z)$$

for all $\theta \in (0,1)$, $y, z \in \mathbb{R}^d$ such that $y - z \in \Lambda$, where

$$\Lambda := \cup_{w \in S^{N-1}} \ker \mathbb{A}(w).$$

Remark 3.5.

(i) The *characteristic cone* Λ as defined in Proposition 3.4 was introduced in the work of Murat and Tartar (see [34, 41]).

(ii) There are $\mathcal{A}$-quasi-convex functions that are not continuous. Indeed, in the degenerate case $\ker \mathcal{A} = \{0\}$ all functions are $\mathcal{A}$-quasi-convex. Furthermore, in general, $Q_{\mathcal{A}}f$ need not be continuous in directions that are not in spanΛ even when f is smooth. As an example, let $N = 1$, $d = 2$, and $\mathcal{A}u := u_2'$. Fix $\varphi \in C^\infty(\mathbb{R})$ such that $0 \le \varphi \le 1$, $\varphi(0) = 1$, $\lim_{|t|\to\infty} \varphi(t) = 0$, and let

$$f(v_1, v_2) := \varphi(v_1 v_2^2).$$

Then $Q_{\mathcal{A}}f$ is obtained by convexification in the first component, and $Q_{\mathcal{A}}f(v_1, v_2) = 0$ if $v_2 \neq 0$, while $Q_{\mathcal{A}}f(v_1, 0) = 1$.

(iii) In the curl-free case and when $\rho = 0$, by Remark 3.3 (iii) we have that $\Lambda = \{a \otimes w : a \in \mathbb{R}^m, w \in S^{N-1}\}$. Thus Proposition 3.4 entails that a quasi-convex Borel measurable function is convex along any rank-one directions. It is then said to be *rank-one convex*. In particular, it is separately convex and so continuous. We remark that although Proposition 3.3 is stated for upper semicontinuous functions f, in the case of gradients the statement still holds if f is only assumed to be Borel measurable (see [21]).

(iv) In the div-free case and by Remark 3.3 (iv), we have $\Lambda = \mathbb{R}^N$, and by Proposition 3.3 we conclude that $Q_{\mathcal{A}}f$ is convex (see also [35]). Thus, since we always have $Q_{\mathcal{A}}f \le f$, $Q_{\mathcal{A}}f$ reduces to the convexification of f.

(v) It follows from the convexity of $t \mapsto Q_{\mathcal{A}}f(a + tz), z \in \Lambda$ (see Proposition 3.4), that $Q_{\mathcal{A}}f(a) > -\infty$ if and only if $Q_{\mathcal{A}}f > -\infty$ on $a + \Lambda$.

Proof of Proposition 3.4.

Case 1. Suppose that f is continuous.

For $R > 0$, $v \in \mathbb{R}^d$, define

$$Q_{\mathcal{A}}^R f(v) := \inf \left\{ \int_{T_N} f(v + w(x))dx \colon w \in C^\infty(T_N) \cap \ker \mathcal{A}, \right.$$
$$\left. \int_{T_N} w(x)dx = 0, \text{ and } ||w||_{L^\infty(T_N)} \le R \right\}.$$

We claim that

$$Q_{\mathcal{A}}^R f \quad \text{is continuous.} \tag{3.2}$$

Let $\rho > 0$, and let ω be the modulus of uniform continuity of f on $B(0, \rho + R)$, i.e.,

$$\omega(r) := \sup\{|f(v) - f(v')| : v, v' \in \overline{B}(0, \rho + R), |v - v'| \le r\}.$$

For all $v, v' \in B(0,\rho)$ and every $w \in C^\infty(T_N) \cap \mathcal{A}$, with $\int_{T_N} w(x)\,dx = 0$ and $||w||_{L^\infty(T_N)} \le R$, we have

$$\begin{aligned}\int_{T_N} f(v + w(x))\,dx &\ge \int_{T_N} f(v' + w(x))\,dx - \omega(|v - v'|)\\ &\ge Q^R_{\mathcal{A}} f(v') - \omega(|v - v'|).\end{aligned}$$

By definition of $Q_{\mathcal{A}} f(v)$ this implies that

$$Q^R_{\mathcal{A}} f(v) - Q^R_{\mathcal{A}} f(v') \ge \omega(|v - v'|)$$

and the uniform continuity of $Q^R_{\mathcal{A}} f$ in $B(0,\rho)$ follows by reversing the roles of v and v'.

Fix $\varepsilon > 0$, let $n \in \mathbb{N}$, and decompose Q into n^N cubes along the coordinate axes, $Q = \cup Q_{n,i}$, $Q_{n,i} = a_{n,i} + (1/n)\,Q$. Now we choose smooth cut-off functions $\varphi_{n,i}$ with the following properties: $0 \le \varphi_{n,i} \le 1$, $\varphi_{n,i} = 1$ on $a_{n,i} + (1/n - 1/n^2)Q$, and $\sum_{i=1}^{n^N} \chi_{Q_{n,i}} \varphi_{n,i} \nearrow 1$. For $w \in C^\infty(T_N) \cap \ker \mathcal{A}$ with average zero on Q, consider the piecewise constant approximations

$$w_n(x) := \sum_{i=1}^{n^N} \chi_{Q_{n,i}}\, w_{n,i}, \quad \text{where } w_{n,i} := n^N \int_{Q_{n,i}} w(x)\,dx.$$

Then $||w_n - w||_{L^\infty(Q)} \to 0$, and by the continuity of $Q^R_{\mathcal{A}} f$ (see (3.2)) we have for $n \ge n_1(\varepsilon)$

$$\begin{aligned}\int_{T_N} Q^R_{\mathcal{A}} f(v + w(x))\,dx &\ge \int_{T_N} Q^R_{\mathcal{A}} f(v + w_n(x))\,dx - \varepsilon\\ &= \sum_{i=1}^{n^N} \frac{1}{n^N}\, Q^R_{\mathcal{A}} f(v + w_{n,i}) - \varepsilon.\end{aligned} \tag{3.3}$$

On the other hand, due to the uniform continuity of f on compact sets, there exists $\delta > 0$ such that

$$\eta,\, \zeta \in L^\infty(\overline{B}(0,5R)), ||\eta - \zeta||_{L^\infty(Q)} < \delta \Rightarrow \left| \int_Q f(v + \eta(x))\,dx - \int_{T_N} f(v + \zeta(x))\,dx \right| < \varepsilon. \tag{3.4}$$

Choose $z_{n,i} \in C^\infty(T_N) \cap \ker \mathcal{A}$, with average zero, such that $||z_{n,i}||_{L^\infty(Q)} \le R$,

$$Q^R_{\mathcal{A}} f(v + w_{n,i}) \ge \int_{T_N} f(v + w_{n,i} + z_{n,i}(y))\,dy - \varepsilon, \tag{3.5}$$

and set

$$y_{n,k}(x) := w(x) + \sum_{i=1}^{n^N} \varphi_{n,i}(x) z_{n,i}(k n^N (x - a_{n,i})), \quad k \in \mathbb{N}.$$

Clearly

$$||y_{n,k}||_{L^\infty(T_N)} \le R + ||w||_{L^\infty(Q)}.$$

By Proposition 2.8, $z_{n,i}(kn^N(\cdot - a_{n,i})) \overset{*}{\rightharpoonup} 0$ in $L^\infty(Q_{n,i})$ as $k \to \infty$, for all $n \in \mathbb{N}$, $i = 1, \dots, n^N$, and so

$$\lim_{k\to\infty} \mathcal{A}y_{n,k} = 0 \quad \text{weak-* in } L^\infty(T_N), \quad \lim_{n\to\infty} \lim_{k\to\infty} \int_{T_N} y_{n,k}\, dx = 0. \tag{3.6}$$

Choose $n = n_2(\varepsilon) \geq n_1(\varepsilon)$ such that

$$n_2(\varepsilon) \to \infty \text{ as } \varepsilon \to 0, \quad ||w_n - w||_{L^\infty(Q)} < \delta, \quad \lim_{k\to\infty} \left| \int_{T_N} y_{n,k}\, dx \right| < \delta. \tag{3.7}$$

Now (3.3), (3.4), (3.5), and (3.7) yield

$$\begin{aligned} \int_{T_N} Q^R_{\mathcal{A}} f(v + w(x))\, dx & \\ &\geq \lim_{k\to\infty} \sum_{i=1}^{n^N} \int_{Q_{n,i}} f(v + w_{n,i} + z_{n,i}(kn^N(x - a_{n,i})))\, dx - 2\varepsilon \\ &\geq \limsup_{k\to\infty} \int_{T_N} f(v + y_{n,k}(x))\, dx - 3\varepsilon \\ &\quad - Cn^N \left[\frac{1}{n^N} - \left(\frac{1}{n} - \frac{1}{n^2} \right)^N \right] \max\{|f(z)| : z \in \overline{B}(v, 2R)\}. \end{aligned} \tag{3.8}$$

In view of Lemma 2.17 and (3.6) we may find $u_k \in L^\infty(T_N; \mathbb{R}^d) \cap \ker \mathcal{A}$ such that

$$\int_{T_N} u_k\, dx = 0, \quad u_k - \left(y_{n,k} - \int_{T_N} y_{n,k}(y)\, dy \right) \to 0 \quad \text{uniformly as} \quad k \to \infty.$$

Thus, by (3.4), (3.7), (3.8), and Remark 3.3 (ii) we have

$$\begin{aligned} \int_{T_N} Q^R_{\mathcal{A}} f(v + w(x))\, dx &\geq \limsup_{k\to\infty} \int_{T_N} f(v + y_{n,k}(x))\, dx - 3\varepsilon - O\left(\frac{1}{n}\right) \\ &\geq \limsup_{k\to\infty} \int_{T_N} f\left(v + y_{n,k}(x) - \int_{T_N} y_{n,k}(y)\, dy \right) dx - 4\varepsilon - O\left(\frac{1}{n}\right) \\ &\geq \limsup_{k\to\infty} \int_{T_N} f(v + u_k(x))\, dx - 5\varepsilon - O\left(\frac{1}{n}\right) \\ &\geq Q_{\mathcal{A}} f(v) - 5\varepsilon - O\left(\frac{1}{n}\right). \end{aligned} \tag{3.9}$$

For $\varepsilon \to 0$ we have, by (3.7), $n = n_2(\varepsilon) \to +\infty$. Hence taking first the limit $\varepsilon \to 0$ and then $R \to \infty$ in (3.9) and observing that $Q^R_{\mathcal{A}} f \searrow Q_{\mathcal{A}} f$ as $R \to \infty$, we deduce from Lebesgue's monotone convergence theorem that

$$\int_{T_N} Q_{\mathcal{A}} f(v + w(x))\, dx \geq Q_{\mathcal{A}} f(v).$$

Case 2. f is upper semicontinuous.

Let $\{f_n\}$ be a sequence of continuous functions converging decreasingly to f. By Case 1, given $v \in \mathbb{R}^d$, $w \in C^\infty(T_N) \cap \ker \mathcal{A}$, with $\int_{T_N} w\, dx = 0$, we have

$$\int_{T_N} Q_{\mathcal{A}} f_n(v + w(x))\, dx \geq Q_{\mathcal{A}} f_n(v), \quad n \in \mathbb{N}.$$

In view of Lebesgue's monotone convergence theorem, $\mathcal{A}$-quasiconvexity of $Q_{\mathcal{A}}f$ will follow provided we show that

$$Q_{\mathcal{A}}f_n \searrow Q_{\mathcal{A}}f. \tag{3.10}$$

Cleary $\{Q_{\mathcal{A}}f_n\}_{n\in\mathbb{N}}$ is decreasing and larger than $Q_{\mathcal{A}}f$. On the other hand, for fixed $v \in \mathbb{R}^d$ with $Q_{\mathcal{A}}f(v) > -\infty$, given $\delta > 0$ there exists $\eta \in C^\infty(T_N) \cap \ker \mathcal{A}$, with $\int_{T_N} \eta\, dx = 0$, such that

$$Q_{\mathcal{A}}f(v) \geq \int_{T_N} f(v + \eta(x))\, dx - \delta.$$

By Lebesgue's monotone convergence theorem it follows that

$$\begin{aligned} Q_{\mathcal{A}}f(v) &\geq \lim_{n\to\infty} \int_{T_N} f_n(v + \eta(x))\, dx - \delta \\ &\geq \limsup_{n\to\infty} Q_{\mathcal{A}}f_n(v) - \delta. \end{aligned}$$

It suffices to let $\delta \to 0$. The case where $Q_{\mathcal{A}}f(v) = -\infty$ is treated in a similar way. As proven in Case 1, the functions $Q_{\mathcal{A}}f_n$ are upper semicontinuous, so $Q_{\mathcal{A}}f = \inf_{n\in\mathbb{N}} Q_{\mathcal{A}}f_n$ is also upper semicontinuous.

Finally, we show that $Q_{\mathcal{A}}f$ is convex on the cones $a + \Lambda$, $a \in \mathbb{R}^d$, i.e.,

$$Q_{\mathcal{A}}f(\theta y + (1-\theta)z) \leq \theta Q_{\mathcal{A}}f(y) + (1-\theta)Q_{\mathcal{A}}f(z)$$

for all $\theta \in (0,1)$, $y, z \in \mathbb{R}^d$ such that $y - z \in \Lambda$. By (3.10) it suffices to prove this inequality in the case where f is a continuous function.

Let

$$\chi(t) := \begin{cases} -(1-\theta) & \text{if} \quad 0 < t < \theta, \\ \theta & \text{if} \quad \theta < t < 1 \end{cases}$$

and extend χ periodically to $\mathbb{R}$ with period one. Let $w \in S^{N-1}$ be such that $y - z \in \ker \mathbb{A}(w)$ and define

$$u_n(x) := (z - y)\, \chi(nx \cdot w).$$

Clearly $u_n \overset{*}{\rightharpoonup} 0$ in $L^\infty(Q)$, and if $\varphi \in C_0^\infty(Q; [0,1])$ is such that $\mathcal{L}^N(\{\varphi = 1\}) = 1 - \delta$, $\delta > 0$, then

$$\mathcal{A}(\varphi u_n) = \sum_{i=1}^N A^{(i)} u_n \frac{\partial \varphi}{\partial x_i} \overset{*}{\rightharpoonup} 0 \qquad \text{in } L^\infty(T_N).$$

Due to Lemma 2.17 we may find $\overline{u}_n \in L^\infty(T_N; \mathbb{R}^d) \cap \ker \mathcal{A}$ such that

$$\int_{T_N} \overline{u}_n = 0, \quad ||\overline{u}_n - \varphi\, u_n||_{L^\infty(Q)} \to 0.$$

By Remark 3.3 (ii), since $Q_{\mathcal{A}}f$ is a $\mathcal{A}$-quasi-convex function, upper semicontinuous, and bounded above by the locally bounded function f, by (3.2), and if $R > 0$ is large

enough, we have

$$\begin{aligned} Q_{\mathcal{A}}f(\theta y+(1-\theta)z) &\le \liminf_{n\to\infty}\int_{T_N} Q_{\mathcal{A}}f(\theta y+(1-\theta)z+\overline{u}_n)\,dx \\ &\le \liminf_{n\to\infty}\int_{T_N} Q^R_{\mathcal{A}}f(\theta y+(1-\theta)z+\overline{u}_n)\,dx \\ &\le \liminf_{n\to\infty}\int_{T_N} Q^R_{\mathcal{A}}f(\theta y+(1-\theta)z+u_n)\,dx + M\delta \\ &= \theta Q^R_{\mathcal{A}}f(\theta y+(1-\theta)z-(1-\theta)(z-y)) \\ &\quad +(1-\theta)Q^R_{\mathcal{A}}f(\theta y+(1-\theta)z+(z-y)\theta)+M\delta \\ &= \theta Q^R_{\mathcal{A}}f(y)+(1-\theta)Q^R_{\mathcal{A}}f(z)+M\delta, \end{aligned}$$

where $M := \max\{|f(z)| : z \in \overline{B}(0,R)\}$. It suffices for us to let $\delta \searrow 0$ and then $R \to \infty$. □

Next we prove that $\mathcal{A}$-quasiconvexity is a necessary condition for lower semicontinuity under the PDE constraint $\mathcal{A}u = 0$.

THEOREM 3.6 (necessity). *Let $f : \Omega \times \mathbb{R}^d \to \mathbb{R}$ be a Carathéodory function such that*

$$\int_\Omega f(x,v(x))\,dx \le \liminf_{n\to\infty}\int_\Omega f(x,v_n(x))\,dx$$

for all sequences $\{v_n\} \subset C^\infty(\overline{\Omega};\mathbb{R}^d)$ that satisfy

$$v_n \overset{*}{\rightharpoonup} v \text{ in } L^\infty(\Omega) \quad \text{and } \mathcal{A}v_n = 0.$$

Assume further that

$$\{f(\cdot,u_n)\} \text{ is equi-integrable}$$

whenever $\{u_n\}$ is a sequence bounded in $L^\infty(\Omega;\mathbb{R}^d)$. Then $f(x_0,\cdot)$ is $\mathcal{A}$-quasi-convex for a.e. $x_0 \in \Omega$.

Proof. Without loss of generality and using a rescaling argument, we may assume that $\Omega \subset Q$.

By the Scorza–Dragoni theorem, for all $i \in \mathbb{N}$ there exists a compact set $K_i \subset \Omega$ such that the restriction of f to $K_i \times \mathbb{R}^d$ is continuous and $\mathcal{L}^N(\Omega \setminus K_i) < 1/i$. Let $\mathcal{S}$ be a countable, dense subset (with respect to uniform convergence) of $\mathbb{W} := \{w \in C^\infty(T_N) : \mathcal{A}w = 0,\ \int_{T_N} w\,dx = 0\}$. Let $x_0 \in \Omega$ be a Lebesgue point for

$$x \mapsto f(x,v), \quad x \mapsto \int_Q f(x,v+w(y))\,dy$$

for all $v \in \mathbb{Q}^d$, $w \in \mathcal{S}$, and suppose that $z \mapsto f(x_0,z)$ is continuous. Fix $v \in \mathbb{Q}^N$, $w \in \mathcal{S}$. We claim that

$$f(x_0,v) \le \int_Q f(x_0,v+w(x))\,dx.$$

If so, by continuity of $z \mapsto f(x_0,z)$ this inequality still holds true for all $v \in \mathbb{R}^d$ and all $w \in \mathbb{W}$. To establish the inequality extend w to $\mathbb{R}^d$ periodically with period Q, fix $\varepsilon > 0$, $h \in \mathbb{N}$, and choose $i = i(h,\varepsilon) \in \mathbb{N}$ such that

$$\mathcal{L}^N\left(Q\left(x_0,\frac{1}{h}\right) \setminus K_i\right) < \frac{\varepsilon}{h^N}.$$

Let $n = n(h, \varepsilon)$ be such that

$$|x - x'| < \frac{1}{n},\ x, x' \in K_i,\ z \in \overline{B}\left(0, |v| + ||w||_{L^\infty(Q)}\right) \quad \Rightarrow |f(x,z) - f(x',z)| < \varepsilon.$$

Decompose the cube $Q\left(x_0, \frac{1}{h}\right)$ as $\cup_{j=1}^{n^N} Q\left(x_j, \frac{1}{hn}\right)$ and if $K_i \cap Q\left(x_j, \frac{1}{hn}\right) \neq \emptyset$ select a_j in this intersection. Choose a cut-off function $\varphi \in C_0^\infty(Q(x_0, 1/h))$ such that $\mathcal{L}^N(Q(x_0, 1/h) \cap \{\varphi \neq 1\}) < \frac{\varepsilon}{h^N}$.

Define

$$w_m(x) := \begin{cases} \varphi(x)\, w^*(hmn(x - x_j)) & \text{if } x \in Q\left(x_j, \frac{1}{hn}\right), j = 1, \dots, n^N, \\ 0, & x \in \mathbb{R}^N \setminus Q\left(x_0, \frac{1}{h}\right), \end{cases}$$

where $w^*(y) := w(y + (1/2, \dots, 1/2))$ for $y \in Q^*$. By Proposition 2.8 it is clear that

$$w_m \overset{*}{\rightharpoonup} 0 \text{ in } L^\infty(T_N), \quad \mathcal{A}w_m \overset{*}{\rightharpoonup} 0 \text{ in } L^\infty(T_N).$$

Using Lemma 2.17 we may find $\eta_m \in L^\infty(Q(0, L); \mathbb{R}^d) \cap \ker \mathcal{A}$ such that $||\eta_m - w_m||_{L^\infty(\Omega)} \to 0$, and so

$$\begin{aligned} \int_\Omega f(x, v)\, dx &\leq \liminf_{m\to\infty} \int_\Omega f(x, v + \eta_m(x))\, dx \\ &= \liminf_{m\to\infty} \int_\Omega f(x, v + w_m(x))\, dx, \end{aligned}$$

where we used Propositions 2.4 and 2.8 and Theorem 2.2 (vi). Taking into account the estimates for $\{\varphi \neq 1\}$ and $Q(x_0, 1/h) \setminus K_i$, we deduce that

$$\begin{aligned} &\int_{Q(x_0,1/h)} f(x, v)\, dx \leq \liminf_{m\to\infty} \int_{Q(x_0,1/h)} f(x, v + w_m(x))\, dx \\ &\quad \leq \liminf_{m\to\infty} \left\{ \sum_{j=1}^{n^N} \int_{Q\left(x_j, \frac{1}{hn}\right) \cap K_i} f(a_j, v + w^*(hmn(x - x_j)))\, dx \right. \\ &\qquad + \sum_{j=1}^{n^N} \int_{Q\left(x_j, \frac{1}{hn}\right) \cap K_i} |f(x, v + w^*(hmn(x - x_j))) - f(a_j, v + w^*(hmn(x - x_j)))|\, dx \\ &\qquad \left. + \sum_{j=1}^{n^N} \int_{Q\left(x_j, \frac{1}{hn}\right) \setminus K_i} f(x, v + w^*(hmn(x - x_j)))\, dx \right\} + M\, \frac{\varepsilon}{h^N} \\ &\quad \leq \sum_{j=1}^{n^N} \frac{1}{(hn)^N} \int_Q f(a_j, v + w(y))\, dy + 3M\, \frac{\varepsilon}{h^N} + \frac{\varepsilon}{h^N}, \end{aligned}$$

where $M := \text{ess sup}\left\{|f(x,z)| : x \in B(x_0, R_0) \subset\subset \Omega, |z| \leq |v| + ||w||_{L^\infty(T_N)}\right\}$.

Hence

$$\begin{aligned} \int_{Q(x_0,1/h)} f(x, v)\, dx &\leq \sum_{j=1}^{n^N} \int_{Q\left(x_j, \frac{1}{hn}\right) \cap K_i} \int_Q f(a_j, v + w(y))\, dy\, dx + \frac{O(\varepsilon)}{h^N} \\ &\leq \sum_{j=1}^{n^N} \int_{Q\left(x_j, \frac{1}{hn}\right)} \int_Q f(x, v + w(y))\, dy\, dx + \frac{O(\varepsilon)}{h^N} \\ &= \int_{Q\left(x_0, \frac{1}{h}\right)} \int_Q f(x, v + w(y))\, dy\, dx + \frac{O(\varepsilon)}{h^N}. \end{aligned} \tag{3.11}$$

Multiplying through (3.11) by h^N, letting $h \to +\infty$, and then $\varepsilon \to 0$, we conclude that

$$f(x_0, v) \le \int_Q f(x_0, v + w(y))\, dy. \qquad \square$$

Now we prove sufficiency of the $\mathcal{A}$-quasiconvexity property.

THEOREM 3.7 (sufficiency). *Let $1 \le p \le +\infty$ and suppose that $f : \Omega \times \mathbb{R}^m \times \mathbb{R}^d \to [0, +\infty)$ is a normal integrand such that $z \mapsto f(x, u, z)$ is $\mathcal{A}$-quasi-convex and continuous for a.e. $x \in \Omega$ and for all $u \in \mathbb{R}^d$. If $1 \le p < +\infty$, then assume further that there exists a locally bounded function $a : \Omega \times \mathbb{R}^d \to [0, +\infty)$ such that*

$$0 \le f(x, u, v) \le a(x, u)(1 + |v|^p).$$

If

$$u_n \to u \text{ in measure}$$

and

$$v_n \rightharpoonup v \text{ in } L^p(\Omega; \mathbb{R}^d) (\overset{*}{\rightharpoonup} \text{ if } p = +\infty),\ \mathcal{A}v_n \to 0 \text{ in } W^{-1,p}(\Omega)\, (\mathcal{A}v_n = 0 \text{ if } p = +\infty),$$

then

$$I(u, v) \le \liminf_{n \to \infty} I(u_n, v_n).$$

This theorem is a consequence of Propositions 3.8 and 3.9.

PROPOSITION 3.8. *Let $1 \le p < +\infty$, and let $\{v_n\}$ be a p-equi-integrable sequence in $L^p(\Omega; \mathbb{R}^d)$ such that $\mathcal{A}v_n \to 0$ in $W^{-1,p}(\Omega)$ if $1 < p < +\infty$, $\mathcal{A}v_n \to 0$ in $W^{-1,r}(\Omega)$ for some $r \in (1, N/(N-1))$ if $p = 1$, and $\{v_n\}$ generates the Young measure $\nu = \{\nu_x\}_{x \in \Omega}$. Let $v_n \rightharpoonup v$ in $L^p(\Omega; \mathbb{R}^d)$. Then for a.e. $a \in \Omega$ there exists a sequence $\{\overline{v}_n\} \subset L^p(T_N; \mathbb{R}^d) \cap \ker \mathcal{A}$ that is p-equi-integrable, generates the homogeneous Young measure ν_a, and satisfies*

$$\int_{T_N} \overline{v}_n\, dx = \langle \nu_a, \mathrm{id} \rangle = v(a).$$

In particular, one has

$$\langle \nu_a, f \rangle \ge f(\langle \nu_a, \mathrm{id} \rangle) = f(v(a))$$

for a.e. $a \in \Omega$ and for every continuous $\mathcal{A}$-quasi-convex f that satisfies

$$|f(z)| \le C(1 + |z|^p)$$

for some $C > 0$ and all $z \in \mathbb{R}^d$.

PROPOSITION 3.9. *Let $\{v_n\}$ be a bounded sequence in $L^\infty(\Omega; \mathbb{R}^d)$ that generates a Young measure $\nu = \{\nu_x\}_{x \in \Omega}$ and satisfies $\mathcal{A}v_n = 0$. Let $v_n \overset{*}{\rightharpoonup} v$ in $L^\infty(\Omega_N; \mathbb{R}^d)$. Then for* a.e. *$a \in \Omega$ and every subcube $Q' \subset\subset Q$, there exists a sequence $\{\overline{v}_n\} \subset L^\infty(T_N; \mathbb{R}^d)$ such that*

$$\overline{v}_n \overset{*}{\rightharpoonup} v(a) \text{ in } L^\infty(T_N),\ \mathcal{A}\overline{v}_n = 0,\ \int_{T_N} \overline{v}_n\, dx = \langle \nu_a, \mathrm{id} \rangle = v(a),$$

and $\{\overline{v}_n\}$ *generates a Young measure* μ *such that*

$$\left|\int_Q \psi(x)\,\langle \mu_x, g\rangle\,dx - \langle \nu_a, g\rangle \int_Q \psi(x)\,dx\right| \le ||g||_{L^\infty(B(0,3M))} \int_{Q\setminus Q'} |\psi(x)|\,dx$$

for all $\psi \in L^1(Q)$, $g \in C_0(\mathbb{R}^d)$, *and where* $M := \sup_{n\in\mathbb{N}} ||v_n||_{L^\infty(\Omega)}$. *In addition, if* $f : \mathbb{R}^d \to \mathbb{R}$ *is a continuous function, then*

$$\langle \nu_a, f\rangle \ge f(\langle \nu_a, \mathrm{id}\rangle) = f(v(a))$$

for a.e. $a \in \Omega$.

We leave the proofs of Propositions 3.8 and 3.9 to the end of this section, and we proceed with the proof of Theorem 3.7. We follow the argument of Kristensen (based on Balder's [4] reasoning for the case without constraints) in the context of the usual curl-free $\mathcal{A}$-quasiconvexity.

Proof of Theorem 3.7. Upon extracting a subsequence, we may assume that

$$\liminf_{n\to\infty} I(u_n, v_n) = \lim_{n\to\infty} I(u_n, v_n),$$

and $\{v_n\}$ generates a Young measure ν. By Proposition 2.5 the pair $\{(u_n, v_n)\}$ generates the Young measure $\{\mu_x = \delta_{u(x)} \otimes \nu_x\}_{x\in\Omega}$, and by Theorem 2.2 (v) we have

$$\begin{aligned}\lim_{n\to\infty} I(u_n, v_n) &\ge \int_\Omega \int_{\mathbb{R}^m\times\mathbb{R}^d} f(x,\eta,\xi)\,d\mu_x(\eta,\xi)\,dx\\ &= \int_\Omega \int_{\mathbb{R}^d} f(x, u(x), \xi)\,d\nu_x(\xi)\,dx.\end{aligned}$$

If $p = 1$ or $p = +\infty$ the result follows from direct application of Proposition 3.8 and Proposition 3.9, respectively, to the map $\xi \mapsto f(x, u(x), \xi)$ and integration over Ω. If $1 < p < +\infty$ then by Lemma 2.15 and by Proposition 2.4, there exists a p-equi-integrable sequence $\{y_n\}$ which generates ν and satisfies $\mathcal{A}y_n = 0$. Once again, it suffices to apply Proposition 3.8 to $\{y_n\}$ and to the map $\xi \mapsto f(x, u(x), \xi)$ for a.e. $x \in \Omega$ fixed. □

Proof of Proposition 3.8. Let $\mathcal{E}$ and $\mathcal{C}$ be countable dense subsets of $L^1(Q)$ and $C_0(\mathbb{R}^d)$, respectively. By Theorem 2.2 (vi) we have

$$g \circ v_n \overset{*}{\rightharpoonup} \langle \nu, g\rangle \text{ in } L^\infty(\Omega)$$

for all $g \in \mathcal{C}$. Let Ω_0 be the set of points $a \in \Omega$ which are Lebesgue points for v, for the functions

$$x \mapsto \int_{\mathbb{R}^d} |\xi|^p\,d\nu_x(\xi), \quad x \mapsto \langle \nu_a, \mathrm{id}\rangle,$$

and for all functions $x \mapsto \langle \nu_x, g\rangle$, $g \in \mathcal{C}$, in the sense that

$$\lim_{R\to 0} \int_Q |\langle \nu_{a+Rx}, g\rangle - \langle \nu_a, g\rangle|\,dx = 0.$$

Consider an increasing sequence of smooth cut-off functions $\varphi_j \in C_0^\infty(Q)$, $\varphi_j \nearrow 1$. For fixed $a \in \Omega_0$, $R > 0$, we define

$$v_{j,R,n}(z) := \varphi_j(z)(v_n(a + Rz) - \langle \nu_a, \mathrm{id}\rangle),\ z \in Q.$$

Recall that $\langle \nu_a, \mathrm{id}\rangle = v(a)$. We have $v_{j,R,n} \in L^p(T_N;\mathbb{R}^d)$, and for all $\psi \in \mathcal{E}$ and $g \in \mathcal{C}$ we have

$$\begin{aligned}\lim_{j\to\infty}\lim_{R\to 0}\lim_{n\to\infty}\int_Q \psi(z)g(v_{j,R,n}(z)+v(a))\,dz &= \lim_{R\to 0}\lim_{n\to\infty}\int_Q \psi(z)\, g(v_n(a+Rz))\,dz\\ &= \lim_{R\to 0}\int_Q \psi(z)\langle \nu_{a+Rz}, g\rangle\,dz\\ &= \langle \nu_a, g\rangle \int_Q \psi(z)\,dz.\end{aligned} \tag{3.12}$$

Moreover, as $\{|v_n|^p\}$ is equi-integrable,

$$\begin{aligned}\limsup_{j\to\infty}\limsup_{R\to 0}\limsup_{n\to\infty}\int_Q |v_{j,R,n}(z)+v(a)|^p\,dz &\le \lim_{R\to 0}\lim_{n\to\infty}\int_Q |v_n(a+Rz)|^p\,dz\\ &= \int_{\mathbb{R}^d}|\xi|^p\,d\nu_a(\xi).\end{aligned} \tag{3.13}$$

Also, $v_{j,R,n} \rightharpoonup 0$ in L^p as $n\to\infty$ and $R\to 0$. If $1 < p < +\infty$ we have, in view of the compact imbedding $L^p(T_N) \hookrightarrow W^{-1,p}(T_N)$ and the assumption $\mathcal{A}v_n \to 0$ in $W^{-1,p}(\Omega)$,

$$\lim_{j\to\infty}\lim_{R\to 0}\lim_{n\to\infty}\mathcal{A}v_{j,R,n} = 0 \quad \text{in } W^{-1,p}(T_N). \tag{3.14}$$

If $p = 1$ then

$$v_{j,R,n} \to 0 \text{ in } W^{-1,r}(T_N) \quad \text{for } r \in \left(1, \frac{N}{N-1}\right),$$

and so, due to (3.12), (3.13), (3.14), and by means of a diagonalization procedure, we may find a sequence of functions $\{w_j\}$ with the properties

$$w_j \rightharpoonup 0 \ \text{ in } L^p(T_N), \quad \mathcal{A}w_j \to 0 \quad \text{in } W^{-1,q}(T_N),$$

where $q = p$ if $1 < p < +\infty$ and $q = r$ if $p = 1$, and

$$\begin{aligned}&\lim_{j\to\infty}\int_Q |w_j(x)+v(a)|^p\,dx = \int_{\mathbb{R}^d}|\xi|^p\,d\nu_a(\xi),\\ &\lim_{j\to\infty}\int_Q \psi(x)g(w_j(x)+v(a))\,dx = \langle \nu_a, g\rangle\int_Q \psi(x)\,dx\end{aligned} \tag{3.15}$$

for all $\psi \in \mathcal{E}$ and $g \in \mathcal{C}$. By Lemmas 2.15 and 2.16 and by (3.15) we conclude that ν_a is generated by a p-equi-integrable sequence $\overline{w}_j \in L^p(T_N;\mathbb{R}^d) \cap \ker \mathcal{A}$ such that $\int_{T_N}\overline{w}_j\,dx = v(a)$. Finally, if f is a continuous function such that $|f(z)| \le C(1+|z|^p)$ for some $C > 0$ and all $z \in \mathbb{R}^d$, then $\{f(\overline{w}_j)\}$ is equi-integrable and by Theorem 2.2 (vi) we have

$$\langle \nu_a, f\rangle = \lim_{j\to\infty}\int_{T_N} f(\overline{w}_j)\,dx \ge f(v(a)),$$

where in the last inequality we used the $\mathcal{A}$-quasiconvexity of f together with Remark 3.2 (ii). ∎

Proof of Proposition 3.9. As in the previous proof, let $\mathcal{E}$ and $\mathcal{C}$ be countable dense subsets of $L^1(Q)$ and $C_0(\mathbb{R}^d)$, respectively, and let Ω_0 be the set of points $a \in \Omega$ which are Lebesgue points for $x \mapsto \langle \nu_x, \mathrm{id}\rangle$ and for all functions $x \mapsto \langle \nu_x, g\rangle$, $g \in \mathcal{C}$. Fix $Q' \subset\subset Q$ and consider a smooth cut-off function $\varphi \in C_0^\infty(Q)$, $0 \le \varphi \le 1$, $\varphi = 1$ in Q'.

For $a \in \Omega_0$, $R > 0$, we define

$$v_{R,n}(z) := \varphi(z)\,(v_n(a + Rz) - \langle \nu_a, \mathrm{id}\rangle) + \langle \nu_a, \mathrm{id}\rangle,\ z \in Q.$$

Then $v_{R,n}$ is bounded in $L^\infty(T_N; \mathbb{R}^d)$, and for all $\psi \in \mathcal{E}$ and $g \in \mathcal{C}$ we have

$$\begin{aligned}\lim_{R\to 0}\lim_{n\to\infty}\int_Q \psi(z) g(v_{R,n}(z))\,dz &= \lim_{R\to 0}\lim_{n\to\infty}\int_Q \psi(z)\, g(v_n(a+Rz))\,dz + \mathcal{E}(\psi, g)\\ &= \lim_{R\to 0}\int_Q \psi(z)\langle \nu_{a+Rz}, g\rangle\,dz + \mathcal{E}(\psi, g)\\ &= \langle \nu_a, g\rangle \int_Q \psi(z)\,dz + \mathcal{E}(\psi, g),\end{aligned}$$

where

$$|\mathcal{E}(\psi, g)| \le ||g||_{L^\infty(B(0,3M))}\int_{Q\setminus Q'} |\psi|\,dy.$$

Clearly, $v_n(a + R\cdot) - \langle \nu_a, \mathrm{id}\rangle \overset{*}{\rightharpoonup} 0$ in L^∞ as $n \to \infty$ and $R \to 0$, and

$$\lim_{R\to 0}\lim_{n\to\infty} \mathcal{A}v_{R,n} = 0 \quad \text{weakly-* in } L^\infty(T_N), \quad \sup_{R,n} ||\mathcal{A}v_{R,n}||_{L^\infty(T_N)} < +\infty.$$

Diagonalizing $\{v_{R,n}\}$, and extracting a further subsequence if necessary, we may find a sequence of functions $\{w_j\}$ with the properties

$$w_j \overset{*}{\rightharpoonup} v(a) \text{ in } L^\infty(T_N), \quad \mathcal{A}w_j \overset{*}{\rightharpoonup} 0 \quad \text{in } L^\infty(T_N),$$

and $\{w_j\}$ generates a Young measure μ such that ess supp$\mu_x \subset B(0, 3M)$ and

$$\left|\int_Q \psi(x)\,\langle \mu_x, g\rangle\,dx - \langle \nu_a, g\rangle \int_Q \psi(x)\,dx\right| \le |\mathcal{E}(\psi, g)|$$

for all $g \in \mathcal{C}, \psi \in \mathcal{E}$. By density this inequality extends to all $\psi \in L^1(Q)$, $g \in C_0(\mathbb{R}^d)$. Due to Lemma 2.17 we may find $\overline{w}_j \in L^\infty(T_N; \mathbb{R}^d) \cap \ker \mathcal{A}$ such that $||w_j - \overline{w}_j||_{L^\infty(T_N)} \to 0$, $\int_{T_N} \overline{w}_j\,dy = v(a)$. In particular, $\{\overline{w}_j\}$ generates the Young measure μ satisfying the statement, and if f is continuous, then

$$\begin{aligned}\lim_{j\to+\infty}\int_{T_N} f(\overline{w}_j)\,dx &= \int_{T_N}\langle \mu_x, f\rangle\,dx\\ &\le \langle \nu_a, f\rangle + \mathcal{L}^N(Q\setminus Q')||f||_{L^\infty(B(0,3R))}.\end{aligned} \tag{3.16}$$

On the other hand, since f is $\mathcal{A}$-quasi-convex and in view of Remark 3.3 (ii) we have directly from Definition 3.1

$$\int_{T_N} f(\overline{w}_j)\,dx \ge f(v(a)) \quad \text{for all } j \in \mathbb{N},$$

which, together with (3.16), and letting $\mathcal{L}^N(Q \setminus Q') \to 0$, concludes the proof. ∎

We end this section with some examples of problems involving PDE constraints which fall within the scope of the present study (for further examples, see [37, 41]).

Example 3.10. (a) Gradients and partial gradients.
The case where

$$\mathcal{A}v = 0 \quad \text{if and only if } v = \nabla u$$

for some function $u : \Omega \to \mathbb{R}^m$ was already treated in Remarks 3.3 (iii) and 3.5 (iii). It can be easily seen that this framework still applies when v is not a full gradient but a list of only some of the partial derivatives of u.

(b) Divergence free fields.
For the example where

$$\mathcal{A}v = 0 \quad \text{if and only if } \mathrm{div} v = 0,$$

we refer the reader to Remarks 3.3 (iv) and 3.5 (iv).

(c) Maxwell's equations.
In magnetostatics the *magnetization* $m : \mathbb{R}^3 \to \mathbb{R}^3$ and the *induced magnetic field* $h : \mathbb{R}^3 \to \mathbb{R}^3$ satisfy (in suitable units) the PDE constraints

$$\mathcal{A}\begin{pmatrix} m \\ h \end{pmatrix} := \begin{pmatrix} \mathrm{div} & (m+h) \\ & \mathrm{curl} h \end{pmatrix} = 0.$$

For $w \in S^2$ we have

$$\begin{aligned} \ker \mathbb{A}(w) &= \left\{(a,b) \in \mathbb{R}^3 \times \mathbb{R}^3 : w \cdot (a+b) = 0, w \otimes b - b \otimes w = 0\right\} \\ &= \left\{(a,b) \in \mathbb{R}^3 \times \mathbb{R}^3 : a \cdot w = -\lambda, b = \lambda w \quad \text{for some } \lambda \in \mathbb{R}\right\}, \end{aligned}$$

and so $\dim \ker \mathbb{A}(w) = 3$ and (2.2) is satisfied. Note also that

$$\Lambda = \left\{(a,b) \in \mathbb{R}^3 \times \mathbb{R}^3 : (a+b) \cdot b = 0\right\},$$

and the fact that Λ imposes no restrictions on a has important consequences in micromagnetics (see [16, 27, 45]). For the full system of Maxwell's equations we refer to [41].

(d) Higher gradients.
Obviously all results remain valid if we replace the target space $\mathbb{R}^d$ by an abstract d-dimensional vector space over $\mathbb{R}$. In order to treat the case of second order derivatives, consider the smooth maps $v : T_N \to E_2^m$, where E_k^m stands for the space of symmetric k-linear maps from $\mathbb{R}^N$ into $\mathbb{R}^m$. Define

$$\mathcal{A}_2 v := \left(\frac{\partial}{\partial x_i} v_{jk} - \frac{\partial}{\partial x_k} v_{ji}\right)_{1 \le i,j,k \le N}.$$

We claim that

$$\left\{v \in C^\infty(T_N; E_2^m) : \mathcal{A}v = 0, \int_{T_N} v \, dx = 0\right\} = \{D^2 u : u \in C^\infty(T_N; \mathbb{R}^m)\}.$$

Indeed, if $\mathcal{A}v = 0$ then $v_{jk} = \frac{\partial w_j}{\partial x_k}$, where $w_j \in C^\infty(\Omega; \mathbb{M}^{m \times N})$ has average zero and is periodic due to the periodicity of v and the fact that $\int_{T_N} v \, dx = 0$. By the

symmetry of v_{jk} we have that curl $w = 0$, and we conclude that $v_{jk} = \frac{\partial^2 u}{\partial x_k \partial x_j}$, where $u \in C^\infty(T_N; \mathbb{R}^m)$.

More generally, in order to study the kth order derivatives of functions $u \in C^\infty(T_N; \mathbb{R}^m)$, we set for $v \in C^\infty(T_N; E_k^m)$

$$\mathcal{A}_k v := \left(\frac{\partial}{\partial x_i} v_{i_1 \dots i_h \, j \, i_{h+2} \dots i_k} - \frac{\partial}{\partial x_j} v_{i_1 \dots i_h \, i \, i_{h+2} \dots i_k} \right)_{0 \le h \le k-1,\ 1 \le i,j,i_1,\dots,i_k \le N}.$$

Here $h = 0$ and $h = k-1$ correspond to the multi-indices $ji_2 \dots i_k$ and $i_1 \dots i_{k-1} j$, respectively. The constant rank condition is satisfied since for $w \in S^{N-1}$

$$\begin{aligned} \ker \mathbb{A}(w) = \{ X \in E_k^m : & \, w_i \, X_{i_1 \dots i_h \, j \, i_{h+2} \dots i_k} - w_j \, X_{i_1 \dots i_h \, i \, i_{h+2} \dots i_k} = 0, \\ & 1 \le h \le k,\ 1 \le i,j,i_1,\dots,i_k \le N \} \\ = \{ X \in E_k^m : & \, X = b \otimes w \dots \otimes w,\ b \in \mathbb{R}^m \} \end{aligned}$$

and so $\dim \ker \mathbb{A}(w) = m$. Moreover,

$$\left\{ v \in C^\infty(T_N; E_k^m) : \mathcal{A} v = 0,\ \int_{T_N} v \, dx = 0 \right\} = \{ D^k u : u \in C^\infty(T_N; \mathbb{R}^m) \}.$$

In fact, if $\mathcal{A} v = 0$, then

$$v_{i_1 \dots i_h j i_{h+2} \dots i_k} = \frac{\partial}{\partial x_j} w_{i_1 \dots i_h i_{h+2} \dots i_k}$$

for some smooth function $w_{i_1 \dots i_h i_{h+1} \dots i_k}$ with average zero. The periodicity of v and the fact that $\int_{T_N} v \, dx = 0$ entail the periodicity of w, and the symmetries of v, together with the zero average condition we imposed on w, imply the symmetry of w, so that $w \in C^\infty(T_N; E_{k-1}^m)$. Furthermore, and once again using the symmetries of v,

$$\begin{aligned} \mathcal{A}_{k-1} w := & \left(\frac{\partial}{\partial x_i} w_{i_1 \dots i_h \, j \, i_{h+2} \dots i_{k-1}} - \frac{\partial}{\partial x_j} w_{i_1 \dots i_h \, i \, i_{h+2} \dots i_{k-1}} \right)_{0 \le h \le k-2,\ 1 \le i,j,i_1,\dots,i_k \le N} \\ & = 0. \end{aligned}$$

The argument may now be completed via induction.

(e) Linear elasticity.

In the framework of linear elasticity, one has to deal with the symmetrized gradient, $v = e(u) := \frac{1}{2}(\nabla u + \nabla^T u)$, of the displacement $u : \Omega \to \mathbb{R}^3$, where $\Omega \subset \mathbb{R}^3$ is an open, bounded set. For $1 < p < +\infty$ one can use a local version of Korn's inequality to reduce the study of functionals

$$u \mapsto I(e(u))$$

to that of functionals

$$u \mapsto J(\nabla u), \quad \text{where } J(\xi) := I\left(\frac{1}{2}(\xi + \xi^T) \right)$$

and proceed as in (a). For $p = 1$ or $p = +\infty$ where one must avoid direct manipulation of the gradient, it is possible to adopt the present framework to treat the second-order operator

$$\tilde{\mathcal{A}} v := \left(\sum_{i=1}^N \frac{\partial^2 v_{ij}}{\partial x_i \partial x_k} + \frac{\partial^2 v_{ik}}{\partial x_i \partial x_j} - \frac{\partial^2 v_{ii}}{\partial x_j \partial x_k} - \frac{\partial^2 v_{jk}}{\partial x_i \partial x_i} \right)_{1 \le j,k \le N}.$$

It turns out that $\tilde{\mathcal{A}}v = 0$ if and only if $v_{ij} = \left(\frac{\partial u_j}{\partial x_i} + \frac{\partial u_i}{\partial x_j}\right)/2$ for some function u. In this setting we have

$$\tilde{\mathcal{A}}v = \sum_{i=1}^{N} A^{(ij)} \frac{\partial^2 v}{\partial x_i \partial x_j}, \quad \tilde{\mathbb{A}}(w) := \sum_{i=1}^{N} A^{(ij)} w_i w_j.$$

(f) Pseudodifferential operators.
The examples (a)–(e) may be treated in a unified way using pseudodifferential operators (see also [44, 45]). For (a)–(d), one considers (on T_N or $\mathbb{R}^N$)

$$\mathcal{B}v := (-\Delta)^{-1/2}\mathcal{A}v = \mathcal{R}_i A^{(i)} v,$$

where $\mathcal{R}_i$ denotes the Riesz transform. For (e) we take

$$\tilde{\mathcal{B}}v := (-\Delta)^{-1}\tilde{\mathcal{A}}u = \left(\sum_{i=1}^{N} \mathcal{R}_i\mathcal{R}_k v_{ij} + \mathcal{R}_i\mathcal{R}_j v_{ik} - \mathcal{R}_j\mathcal{R}_k v_{ii} - v_{jk}\right)_{1\le j,k\le N}. \tag{3.17}$$

The symbol of $\mathcal{B}$ is

$$b(\xi) := \frac{\xi_i}{|\xi|} A^{(i)},$$

and the constant rank condition becomes rank $b(\xi) = r$ for all $\xi \neq 0$. Similarly, for (3.17) the symbol takes values in $\mathrm{Lin}(\mathrm{E}_2, \mathrm{E}_2)$ and is given by

$$\tilde{b}(\xi)M := M\xi \otimes \xi + \xi \otimes M\xi - (\xi \otimes \xi)\,\mathrm{tr}M - M.$$

One can easily check that if $|\xi| = 1$, then

$$\ker \tilde{b}(\xi) = \left\{a \otimes \xi + \xi \otimes a : a \in \mathbb{R}^N\right\},$$

which has dimension N. Hence $\tilde{\mathcal{B}}$ satisfies the analogue of (2.2).

4. Characterization of Young measures. The result below is the generalization to the $\mathcal{A}$-free setting of the theorem by Kinderlehrer and Pedregal for the case of gradients [28, 29]. We roughly follow their strategy that relies on the Hahn–Banach separation theorem and the representation of the ($\mathcal{A}$-)quasi-convex envelope (see (3.1) and Proposition 3.4). Tartar [41] has earlier used the Hahn–Banach separation theorem to characterize Young measures in the case without differential constraints. (In a similar vein, Berliocchi and Lasry [11] used the Krein–Milman theorem.) Our presentation closely follows Kristensen's strategy for the case of gradients. We first establish the result for $p = 1$ and then deduce the assertion for $1 < p < +\infty$ by a truncation process. Some of our arguments are similar to those of Sychev [40] who, independently of our work, proposed an alternative approach to gradient Young measures.

THEOREM 4.1. *Let $1 \le p < +\infty$, and let $\{\nu_x\}_{x\in\Omega}$ be a weakly measurable family of probability measures on $\mathbb{R}^d$. There exists a p-equi-integrable sequence $\{v_n\}$ in $L^p(\Omega;\mathbb{R}^d)$ that generates the Young measure ν and satisfies $\mathcal{A}v_n = 0$ in Ω if and only if the following three conditions hold:*

(i) *there exists $v \in L^p(\Omega;\mathbb{R}^d)$ such that $\mathcal{A}v = 0$ and*

$$v(x) = \langle \nu_x, \mathrm{id}\rangle \quad \text{a.e. } x \in \Omega;$$

(ii)

$$\int_\Omega \int_{\mathbb{R}^d} |z|^p \, d\nu_x(z) \, dx < +\infty;$$

(iii) *for* a.e. $x \in \Omega$ *and all continuous functions* g *that satisfy* $|g(v)| \le C(1+|v|^p)$ *for some* $C > 0$ *and all* $v \in \mathbb{R}^d$ *one has*

$$\langle \nu_x, g \rangle \ge Q_{\mathcal{A}} g(\langle \nu_x, \mathrm{id} \rangle).$$

Remark 4.2.

(i) From Lemma 2.15 it follows that if $1 < p < +\infty$ properties (i)–(iii) are still necessary if the condition $\mathcal{A}v_n = 0$ is replaced by the weaker requirement $\mathcal{A}v_n \to 0$ in $W^{-1,p}(\Omega)$.

(ii) In view of Theorem 2.2 (i) it suffices to assume that $\nu_x \ge 0$ a.e. $x \in \Omega$. Condition (iii) then implies $\nu_x(\mathbb{R}^d) = 1$.

(iii) A similar statement is valid for operators with variable coefficients, as long as rank $\mathbb{A}(x, w)$ is constant for all $w \in S^{N-1}$ and a.e. $x \in \Omega$. Such results are, however, more naturally discussed in the context of pseudodifferential constraints and will appear elsewhere. For the quadratic case, see [45].

Proof of Theorem 4.1—Necessity. Necessity of (i) follows immediately from Theorem 2.2 (vi), where v is the weak limit in L^p of the sequence $\{v_n\}$. Property (ii) is deduced from Theorem 2.2 (v) with $f(z) = |z|^p$, and (iii) is a consequence of Proposition 3.8 (and Lemma 2.15 if $1 < p < +\infty$). □

The proof of sufficiency for $1 < p < +\infty$ follows from the case $p = 1$ and Corollary 2.18.

We proceed with the proof in the case of homogeneous $\mathcal{A}$-1-Young measures.

Let $\mathcal{P}$ be the set of probability measures on $\mathbb{R}^d$ and define

$$\mathbb{H} := \{\nu \in \mathcal{P}(\mathbb{R}^d) : \langle \nu, \mathrm{id} \rangle = 0, \text{ there exists an equi-integrable sequence } \{w_j\} \subset L^1(T_N) \cap \ker \mathcal{A} \text{ generating the Young measure } \nu\}.$$

Set

$$E := \left\{ g \in C(\mathbb{R}^d) : \lim_{|z| \to \infty} \frac{g(z)}{1+|z|} \quad \text{exists in } \mathbb{R} \right\}$$

equipped with the norm

$$||g||_E := \sup_{z \in \mathbb{R}^d} \frac{|g(z)|}{1+|z|}.$$

This space is isometrically isomorphic to the space $C(\mathbb{R}^d \cup \{\infty\}) \sim C(S^d)$ of continuous functions on the one-point compactification of $\mathbb{R}^d$, via the map

$$g \mapsto \frac{g(\cdot)}{1+|\cdot|}.$$

In particular, E is a separable Banach space, and its dual E' may be identified with the space of Radon measures on $\mathbb{R}^d \cup \{\infty\}$. Thus if $\nu \in \mathcal{P}$ is such that

$$\int_{\mathbb{R}^d} |z| \, d\nu(z) < +\infty,$$

then $\nu \in E'$ since for all $g \in E$

$$\left| \int_{\mathbb{R}^d} g \, d\nu \right| \leq ||g||_E \int_{\mathbb{R}^d} (1 + |z|) \, d\nu(z).$$

PROPOSITION 4.3. *Let $\nu \in \mathcal{P}(\mathbb{R}^d)$ with $\langle \nu, \mathrm{id} \rangle = 0$. Then $\nu \in \mathbb{H}$ if*

(i)

$$\int_{\mathbb{R}^d} |z| \, d\nu(z) < +\infty;$$

(ii)

$$\langle \nu, g \rangle \geq Q_{\mathcal{A}} g(0)$$

for all $g \in C(\mathbb{R}^d)$ such that $|g(z)| \leq C(1 + |z|)$.

Proof. We follow [28, 29] and use the Hahn–Banach theorem to show that measures satisfying (i) and (ii) cannot be separated from $\mathbb{H}$.

We will prove that $\mathbb{H}$ is convex and relatively closed in $\mathcal{P}$.

Claim 1. $\mathbb{H}$ is convex.

Fix $\nu, \mu \in \mathbb{H}$, $\theta \in (0,1)$. Let $\{v_j\}, \{w_j\} \subset L^1(T_N) \cap \ker \mathcal{A}$ be equi-integrable sequences generating the $\mathcal{A}$-1-Young measures ν and μ, respectively. By means of a mollification, we may take $v_j, w_j \in C^\infty(T_N)$. Also, as

$$\int_{T_N} v_j \, dx, \ \int_{T_N} w_j \, dx \to 0,$$

without loss of generality we may assume that

$$\int_{T_N} v_j \, dx = \int_{T_N} w_j \, dx = 0.$$

Since $v_j, w_j \to 0$ in $W^{-1,p}(T_N)$ for $p < \frac{N}{N-1}$, and as for all $\varphi \in C_0^\infty((0,\theta) \times T_{N-1})$

$$||\mathcal{A}(\varphi(w_j - v_j))||_{W^{-1,p}} = \left|\left| \frac{\partial \varphi}{\partial x_i} A^{(i)}(w_j - v_j) \right|\right|_{W^{-1,p}} \to 0,$$

we may find a sequence $\{\varphi_j\} \subset C_0^\infty((0,\theta) \times T_{N-1})$ such that $\varphi_j \nearrow \chi_{(0,\theta)\times T_{N-1}}$ and

$$||\mathcal{A}(\varphi_j(w_j - v_j))||_{W^{-1,p}} \to 0.$$

Define

$$u_j := v_j + \mathbb{T}\left(\varphi_j(w_j - v_j) - \int_{T_N} \varphi_j(w_j - v_j) \, dy \right).$$

Then $u_j \in L^1(T_N) \cap \ker \mathcal{A}$, $\int_{T_N} \varphi_j(w_j - v_j) \, dy \to 0$, and by Lemma 2.14 (iii),

$$u_j = v_j + \varphi_j(w_j - v_j) + h_j, \quad h_j \to 0 \text{ in } L^p(T_N), \, p < \frac{N}{N-1}.$$

In particular, $\{u_j\}$ is equi-integrable and generates the Young measure $\{\lambda_x\}_{x \in T_N}$ given by

$$\lambda_x = \begin{cases} \nu & \text{if } x_1 \in (0,\theta), \\ \mu & \text{if } x_1 \in (\theta,1). \end{cases}$$

Finally, let

$$\overline{u}_{j,m}(x) := u_j(mx),\ m \in \mathbb{N}.$$

Then $\overline{u}_{j,m} \in C^\infty(T_N) \cap \ker \mathcal{A}$, by periodicity $\sup_{j,m} ||\overline{u}_{j,m}||_{L^1(T_N)} < +\infty$, and due to the equi-integrability of $\{u_j\}$, for all $\psi \in C_0(\mathbb{R}^N)$, $g \in E$, we have

$$\begin{aligned}\lim_{j\to\infty}\lim_{m\to\infty}\int_{\mathbb{R}^N}\psi(x)g(\overline{u}_{j,m}(x))\,dx &= \lim_{j\to\infty}\int_{\mathbb{R}^N}\psi(x)\left(\int_{T_N} g(u_j(y))\,dy\right)dx \\ &= \int_{\mathbb{R}^N}\psi(x)\,dx\,(\theta\langle\nu,g\rangle + (1-\theta)\langle\mu,g\rangle). \qquad (4.1)\end{aligned}$$

Extracting a diagonal subsequence and taking $g = |\cdot|$ in (4.1), by Theorem 2.2 (vi) we conclude that $\theta\nu + (1-\theta)\mu$ is generated by an equi-integrable sequence in $\ker \mathcal{A}$ and thus belongs to $\mathbb{H}$.

Claim 2. $\mathbb{H}$ is relatively closed in $\mathcal{P}$ with respect to the weak-* topology in E', i.e.,

$$\overline{\mathbb{H}}^{E'} \cap \mathcal{P} = \mathbb{H}.$$

Let $\nu \in \overline{\mathbb{H}}^{E'} \cap \mathcal{P}$, let $\{f_i\}_{i\in\mathbb{N}} \subset C^\infty(T_N)$ be dense in $L^1(T_N)$, and let $\{g_j\}_{j\in\mathbb{N}} \subset C_0^\infty(\mathbb{R}^d)$ be dense in $C_0(\mathbb{R}^d)$. We take $f_0 = 1$ and $g_0(z) = |z|$. By definition of weak-* topology in E' there exist $\nu_k \in \mathbb{H}$ such that

$$|\langle \nu - \nu_k, g_j\rangle| < \frac{1}{2k}, \quad j = 0, \dots, k;$$

thus, by virtue of Theorem 2.2 (vi) we may find $w_k \in L^1(T_N) \cap \ker \mathcal{A}$ such that

$$\left|\langle\nu, g_j\rangle \int_{T_N} f_i\,dx - \int_{T_N} f_i g_j(w_k)\,dx\right| < \frac{1}{k}, \quad 0 \le i,j \le k. \qquad (4.2)$$

In particular, setting $i = 0 = j$ we deduce that $\{w_k\}$ is bounded in $L^1(T_N)$ and so (a subsequence) generates a Young measure μ. From (4.2) and the density properties of $\{f_i\}_{i\in\mathbb{N}}$ and $\{g_j\}_{j\in\mathbb{N}}$ it follows that $\mu = \nu$, and the choice $i = 0 = j$ yields

$$\int_{T_N} |w_k|\,dx \to \langle\nu, |\cdot|\rangle.$$

By Theorem 2.2 (vi) we conclude that $\{w_k\}$ is equi-integrable and so $\nu \in \mathbb{H}$. This proves Claim 2.

Consider $\nu \in \mathcal{P}$ such that $\langle\nu, \mathrm{id}\rangle = 0$ and ν satisfies (i), (ii). We want to prove that $\nu \in \mathbb{H}$. Suppose that $\nu \notin \mathbb{H}$. By Claims 1 and 2, $\nu \notin \overline{\mathrm{co}(\mathbb{H})}$ with respect to the weak-* topology of E'. Therefore, by the Hahn–Banach theorem and (ii) there exist $g \in E$, $\alpha \in \mathbb{R}$, such that

$$\langle\mu, g\rangle \ge \alpha \quad \text{for all } \mu \in \mathbb{H}, \qquad Q_{\mathcal{A}}g(0) \le \langle\nu, g\rangle < \alpha. \qquad (4.3)$$

Given $w \in C^\infty(T_N) \cap \ker \mathcal{A}$, with $\int_{T_N} w\,dx = 0$, by Proposition 2.8 we have $\overline{\delta_w} \in \mathbb{H}$ and thus

$$\int_{T_N} g(w)\,dx = \langle\overline{\delta_w}, g\rangle \ge \alpha,$$

which, by Definition 3.1 implies that $Q_{\mathcal{A}}g(0) \geq \alpha$, contradicting (4.3). We conclude that $\nu \in \mathbb{H}$. □

Next we treat the case of inhomogeneous $\mathcal{A}$-1-Young measures.

We define

$$\mathbb{X} := \Bigg\{ \nu : \Omega \to \mathcal{P}(\mathbb{R}^d) : \nu \text{ is weak* measurable,} \int_\Omega \int_{\mathbb{R}^d} |z| d\nu_x(z) dx < +\infty, \langle \nu_x, \mathrm{id} \rangle = 0 \text{ a.e. } x \in \Omega \Bigg\},$$

$$\mathbb{Y} := \Bigg\{ \nu \in \mathbb{X} : \nu \text{is generated by an equi-integrable sequence} \{w_n\} \in L^1(T_N) \cap \ker \mathcal{A} \Bigg\},$$

$$\mathbb{W} := \{ \nu \in \mathbb{X} : \langle \nu_x, g \rangle \geq Q_{\mathcal{A}} g(0) \text{ a.e. } x \in \Omega \text{ and for all } g \in E \},$$

and

$$\mathcal{E} := C(\overline{\Omega}; E) \sim C(\overline{\Omega} \times (\mathbb{R}^d \cup \{\infty\})).$$

Suppose that ν satisfies (i), (ii), and (iii) of Theorem 4.1, and set $\overline{\nu}_x := \Gamma_{-v(x)} \nu_x$. (The translation of a measure was defined in Proposition 2.4.) Clearly $\overline{\nu} \in \mathbb{W}$, and so if $\mathbb{W} \subset \mathbb{Y}$, then ν is generated by an equi-integrable sequence $\{v + w_j\}$ where $\mathcal{A} w_j = 0$. It thus suffices to verify the following assertion.

PROPOSITION 4.4.

$$\mathbb{W} \subset \mathbb{Y}. \tag{4.4}$$

Proof. The strategy to prove (4.4) is as follows.

Step 1. $\overline{\mathbb{Y}}^{\mathcal{E}'} \cap \mathbb{X} = \mathbb{Y}$ in the weak-* topology.

Step 2. It is possible to find a good subset $D \subset \mathbb{W}$ such that $\overline{D}^{\mathcal{E}'} \cap \mathbb{W} = \mathbb{W}$.

Step 3. $D \subset \mathbb{Y}$.

The proof of Step 1 is entirely identical to that of Claim 2 in the proof of Proposition 4.3. For Step 2, we define $\mathcal{G}_k$ to be the family of cubes of the form

$$\left\{ \frac{1}{k}(y + Q) : y \in \mathbb{Z}^N, \frac{1}{k}(y + Q) \subset \Omega \right\},$$

and we set

$$G_k := \cup_{\mathcal{U} \in \mathcal{G}_k} \mathcal{U}.$$

Consider the sets of piecewise homogeneous Young measures

$$\mathbb{W}_k := \left\{ \nu \in \mathbb{W} : \nu_{|\mathcal{U}} \text{ is homogeneous if } \mathcal{U} \in \mathcal{G}_k,\ \nu_{|(\Omega \setminus G_k)} = \delta_0 \right\},$$

and let

$$D := \cup_{k \in \mathbb{N}} \mathbb{W}_k.$$

In order to show that

$$\overline{D}^{\mathcal{E}'} \cap \mathbb{W} = \mathbb{W},$$

let $\nu \in \mathbb{W}$ and define

$$\nu_x^k := \begin{cases} \frac{1}{\mathcal{L}^N(\mathcal{U})} \int_{\mathcal{U}} \nu_y \, dy & \text{if } x \in \mathcal{U},\ \mathcal{U} \in \mathcal{G}_k, \\ \delta_0 & \text{otherwise.} \end{cases}$$

It is clear that $\nu^k \in \mathbb{W}_k$, so it suffices to show that

$$\langle \nu^k, f \rangle \to \langle \nu, f \rangle \qquad \text{for all } f \in \mathcal{E}. \tag{4.5}$$

Fix $f \in \mathcal{E}$, and for each $\mathcal{U} \in \mathcal{G}_k$ denote by $x_{\mathcal{U}} \in \left(\frac{1}{k}\mathbb{Z}\right)^N$ the lower left corner of $\mathcal{U}$ so that $\mathcal{U} = x_{\mathcal{U}} + \frac{1}{k}Q$. Let ω be a modulus of uniform continuity of f, i.e.,

$$\omega(\delta) := \sup\left\{ ||f(x,\cdot) - f(y,\cdot)||_E : \, x, y \in \overline{\Omega},\ |x-y| \le \delta \right\}.$$

We have

$$\begin{aligned} &\left| \int_{\mathcal{U}} \int_{\mathbb{R}^d} f(x,z) \, d\nu_x(z) dx - \int_{\mathcal{U}} \int_{\mathbb{R}^d} f(x,z) \, d\nu_x^k(z) \, dx \right| \\ &\quad \le \left| \int_{\mathcal{U}} \int_{\mathbb{R}^d} f(x_{\mathcal{U}},z) \, d\nu_x(z) \, dx - \int_{\mathcal{U}} \int_{\mathbb{R}^d} f(x_{\mathcal{U}},z) \, d\nu_x^k(z) \, dx \right| \\ &\qquad + \omega\left(\frac{1}{k}\right) ||f||_{\mathcal{E}} \left(\int_{\mathcal{U}} \int_{\mathbb{R}^d} (1+|z|) \, d\nu_x(z) \, dx + \int_{\mathcal{U}} \int_{\mathbb{R}^d} (1+|z|) \, d\nu_x^k(z) \, dx \right) \\ &\quad \le 2\omega\left(\frac{1}{k}\right) ||f||_{\mathcal{E}} \int_{\mathcal{U}} \int_{\mathbb{R}^d} (1+|z|) \, d\nu_x(z) dx. \end{aligned}$$

Therefore,

$$\begin{aligned} |\langle \nu^k, f \rangle - \langle \nu, f \rangle| \le\ & 2\omega\left(\frac{1}{k}\right) ||f||_{\mathcal{E}} \int_{G_k} \int_{\mathbb{R}^d} (1+|z|) \, d\nu_x(z) dx \\ & + 2||f||_{\mathcal{E}} \int_{\Omega \setminus G_k} \int_{\mathbb{R}^d} (1+|z|) \, d\nu_x(z) dx, \end{aligned}$$

and (4.5) follows by letting $k \to \infty$ and using assertion (ii) in Theorem 4.1.

Next, we carry out Step 3 by showing that

$$\mathbb{W}_k \subset \mathbb{Y} \quad \text{for all } k \in \mathbb{N}.$$

Using a rescaling argument, we may assume that $\Omega \subset Q$. Fix $k \in \mathbb{N}$ and let $\mathcal{G}_k = \{Q_i\}_{i=1}^m$ for some $m \in \mathbb{N}$. Fix $\nu \in \mathbb{W}_k$, with $\nu_{|Q_i} = \nu^i$. By Corollary 2.18 for each $i \in \{1, \dots, m\}$ there exists an equi-integrable sequence $\{w_j^i\} \subset L^1(T_N) \cap \ker \mathcal{A}$ generating ν^i. In particular, without loss of generality we may assume that w_j^i are smooth, and that we have

$$w_j^i \rightharpoonup 0 \quad \text{in } L^1(Q_i), \quad w_j^i \to 0 \quad \text{in } W^{-1,p}_{loc}(\mathbb{R}^N)$$

for $p < N/(N-1)$. Hence, we may find smooth cut-off functions $\varphi_j^i \in C_0^\infty(Q_i; [0,1])$ such that $\varphi_j^i \nearrow \chi_{Q_i}$ and

$$\mathcal{A}\left(\sum_{i=1}^m \varphi_j^i w_j^i \right) = \sum_{k=1}^N \sum_{i=1}^m A^{(k)} w_j^i \frac{\partial \varphi_j^i}{\partial x_k} \ \to 0 \quad \text{in } W^{-1,p}(\mathbb{R}^N).$$

Setting

$$u_j := \mathbb{T}\left(w_j - \int_{T_N} w_j\, dy\right), \quad \text{where } w_j := \sum_{i=1}^{m} \varphi_j^i\, w_j^i,$$

then $u_j \in \ker \mathcal{A}$, $\left|\left|u_j - \sum_{i=1}^m \varphi_j^i\, w_j^i\right|\right|_{L^p(\Omega)} \to 0$. In particular $\{u_j\}$ is equi-integrable and it generates ν, so $\nu \in \mathbb{Y}$. □

Example 4.5. (a) Gradients.

Using Remark 3.3 (iii) and Theorem 4.1, we recover the characterization of $W^{1,p}$ gradient Young measures as obtained by Kinderlehrer and Pedregal [28, 29] (see Theorem 2.6).

(b) Divergence-free fields.

It follows from Remarks 3.3 (iv), 3.5 (iv), and Theorem 4.1 that any weakly measurable family of probability measures $\{\nu_x\}_{x\in\Omega}$ satisfying

$$\operatorname{div}(\langle \nu_x, \mathrm{id}\rangle) = 0, \quad \int_\Omega \int_{\mathbb{R}^N} |z|^p\, d\nu_x(z) dx < +\infty,$$

is generated by a p-equi-integrable sequence of divergence-free fields $v_n \in L^p(\Omega; \mathbb{R}^N)$ (see also [35]).

(c) Micromagnetics.

In view of Example 3.10 c), we may apply Theorem 4.1 to the system of Maxwell equations. Moreover, if $1 < p < +\infty$, if ν is an $\mathcal{A}$-p-Young measure, and if we define the projection λ by

$$\lambda_x(U) := \nu_x(U \times \mathbb{R}^3) \quad \text{for any open subset } U \subset \mathbb{R}^3,$$

then supp $\lambda_x \subset S^2$ for a.e.$x \in \Omega$ if and only if ν is generated by a p-equi-integrable sequence $\{\tilde{m}_n, \tilde{h}_n)\} \subset \ker \mathcal{A}$ such that $|\tilde{m}_n(x)| = 1$ for a.e. $x \in \Omega$. Indeed, assuming that λ_x is supported on the unit sphere, let $\{(m_n, h_n)\} \subset \ker \mathcal{A}$ be a p-equi-integrable generating sequence, with $h_n = -\nabla u_n$, $u_n \in W_0^{1,p}(\Omega)$ ($h_n = -\nabla u_n + H_n$ with div $H_n =$ curl $H_n = 0$ if Ω is not simply connected). Consider the projection

$$\pi(x) := \begin{cases} \frac{x}{|x|} & \text{if } x \neq 0, \\ \\ x_0 & \text{if } x = 0, \end{cases}$$

where $x_0 \in S^2$ is fixed, and define $\tilde{m}_n := \pi m_n$. Since $\operatorname{dist}(m_n, S^2) \to 0$ as $n \to \infty$, we have that $\tilde{m}_n - m_n \to 0$ in measure, and, due to the p-equi-integrability, we conclude that $\tilde{m}_n - m_n \to 0$ in L^p. Let $\tilde{h}_n := -\nabla \tilde{u}_n$ ($\tilde{h}_n := -\nabla \tilde{u}_n + H_n$ if Ω is not simply connected), where $\tilde{u}_n \in W_0^{1,p}(\Omega)$ and $\operatorname{div}(\tilde{m}_n - \nabla \tilde{u}_n) = 0$. We have

$$\operatorname{div}((\tilde{m}_n - m_n) - (\nabla \tilde{u}_n - \nabla u_n)) = 0;$$

therefore $\Delta(\tilde{u}_n - u_n) \to 0$ in $W^{-1,p}$, and thus $\tilde{u}_n - u_n \to 0$ in $W^{1,p}$. We conclude that $\{(\tilde{m}_n, \tilde{h}_n)\}$ still generates ν.

REFERENCES

[1] E. Acerbi and N. Fusco, *Semicontinuity problems in the calculus of variations*, Arch. Rational Mech. Anal., 86 (1984), pp. 125–145.

[2] J. J. Alibert and G. Bouchitté, *Non uniform integrability and generalized Young measures*, J. Convex Anal., 4 (1997), pp. 129–147.

[3] L. Ambrosio and G. Dal Maso, *On the relaxation in $BV(\Omega;\mathbb{R}^m)$ of quasi–convex integrals*, J. Funct. Anal., 109 (1992), pp. 76–97.

[4] E. J. Balder, *A general approach to lower semicontinuity and lower closure in optimal control theory*, SIAM J. Control Optim., 22 (1984), pp. 570–598.

[5] J. M. Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Rational Mech. Anal., 63 (1977), pp. 337–403.

[6] J. M. Ball, *A version of the fundamental theorem for Young measures*, in PDE's and Continuum Models of Phase Transitions, M. Rascle, D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 344, Springer-Verlag, Berlin, 1989, pp. 207–215.

[7] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[8] 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, 338 (1992), pp. 389–450.

[9] J. M. Ball and F. Murat, *$W^{1,p}$-quasiconvexity and variational problems for multiple integrals*, J. Funct. Anal, 58 (1984), pp. 222–253.

[10] J. M. Ball and F. Murat, *Remarks on rank-one convexity and quasiconvexity*, in Ordinary and Partial Differential Equations, Vol. III, B. D. Sleeman and R. J. Jarvis, eds., Longman, Harlow, UK, 1991, pp. 25–37.

[11] H. Berliocchi and J.-M. Lasry, *Intégrands normales et mesures paramétrées en calcul des variations*, Bull. Soc. Math. France, 101 (1973), pp. 129–184.

[12] G. Bouchitté, I. Fonseca, and L. Mascarenhas, *A global method for relaxation*, Arch. Rational Mech. Anal, 145 (1998), pp. 55–98.

[13] M. Chipot and D. Kinderlehrer, *Equilibrium configurations of crystals*, Arch. Rational Mech. Anal., (1998), pp. 237–277.

[14] B. Dacorogna, *Weak Continuity and Weak Lower Semicontinuity for Nonlinear Functionals*, Lecture Notes in Math. 922, Springer-Verlag, New York, 1982.

[15] B. Dacorogna, *Quasiconvexity and relaxation of non convex variational problems*, J. Funct. Anal., 46 (1982), pp. 102–118.

[16] A. De Simone, *Energy minimizers for large ferromagnetic bodies*, Arch. Rational Mech. Anal., 125 (1993), pp. 99–143.

[17] R. J. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Rational Mech. Anal., 82 (1983), pp. 27–70.

[18] R. J. DiPerna, *Compensated compactness and general systems of conservation laws*, Trans. Amer. Math. Soc., 292 (1985), pp. 383–42.

[19] 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.

[20] L. C. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, CBMS Regional Conf. Ser. in Math. 74, AMS, Providence, RI, 1990.

[21] I. Fonseca, *The lower quasiconvex envelope of the stored energy function for an elastic crystal*, J. Math. Pures Appl., 67 (1988), pp. 175–195.

[22] I. Fonseca and S. Müller, *Quasiconvex integrands and lower semicontinuity in L^1*, SIAM J. Math. Anal., 23 (1992), pp. 1081–1098.

[23] I. Fonseca and S. Müller, *Relaxation of quasiconvex functionals in $BV(\Omega,\mathbb{R}^p)$ for integrands $f(x,u,\nabla u)$*, Arch. Rational Mech. Anal., 123 (1993), pp. 1–49.

[24] I. Fonseca, S. Müller, and P. Pedregal, *Analysis of concentration and oscillation effects generated by gradients*, SIAM J. Math. Anal., 29 (1998), pp. 736–756.

[25] P. Gerard, *Compacité par compensation et regularité 2-microlocale*, Séminaire Eq. aux Dér. Part., Ecole Polytechnique, Palaiseau, exp VI, 1988-89.

[26] P. Gerard, *Microlocal defect measure*, Comm. Partial Differential Equations, 16 (1989), pp. 1761–1794.

[27] R. James and D. Kinderlehrer, *Theory of magnetostriction with applications to $Tb_xDy_{1-x}Fe_2$*, Phil. Mag. B, 68 (1993), pp. 237–274.

[28] D. Kinderlehrer and P. Pedregal, *Characterizations of Young measures generated by gradients*, Arch. Rational Mech. Anal., 115 (1991), pp. 329–365.

[29] D. Kinderlehrer and P. Pedregal, *Gradient Young measures generated by sequences in Sobolev spaces*, J. Geom. Anal., 4 (1994), pp. 59–90.

[30] J. Kristensen, *Finite Functionals and Young Measures Generated by Gradients of Sobolev Functions*, Mat-Report 1994-34, Mathematical Institute, Technical University of Denmark, Lyngby, Denmark, 1994.
[31] P. Marcellini, *Approximation of quasiconvex functions and semicontinuity of multiple integrals*, Manuscripta Math., 51 (1985), pp. 1–28.
[32] C. B. Morrey, *Multiple Integrals in the Calculus of Variations*, Springer-Verlag, New York, 1966.
[33] S. Müller, *Rank-One Convexity Implies Quasiconvexity on Diagnol Matrices*, Preprint 23/1999, Max-Planck Institute for Mathematics in the Sciences, Leiptig, Germany, 1999.
[34] F. Murat, *Compacité par compensation : condition necessaire et suffisante de continuité faible sous une hypothése de rang constant*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 8 (1981), pp. 68–102.
[35] P. Pedregal, *Parametrized Measures and Variational Principles*, Birkhäuser, Boston, 1997.
[36] M. E. Schonbeck, *Convergence of solutions to non-linear dispersive equations*, Comm. in Partial Differential Equations, 7 (1982), pp. 959–1000.
[37] J. R. Schulenberger and C. H. Wilcox, *Coerciveness inequalities for nonelliptic systems of partial differential equations*, Annali di Matematica, 88 (1971), pp. 229–305.
[38] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.
[39] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton, NJ, 1971.
[40] M. A. Sychev, *A new approach to Young measure theory, relaxation and convergence in energy*, Ann. IHP Anal. Non Linéaire, Inst. H. Poincaré, to appear.
[41] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, R. Knops, ed., Vol. IV, Pitman Res. Notes Math. 39, Longman, Harlow, UK, 1979, pp. 136–212.
[42] L. Tartar, *The compensated compactness method applied to systems of conservation laws*, in Systems of Nonlinear Partial Differential Equations, J. M. Ball, ed., D. Riedel, Dordrecht, 1983.
[43] L. Tartar, *Étude des oscillations dans les équations aux dérivées partielles nonlinéaires*, in Trends and Applications of Pure Mathematics to Mechanics, Lecture Notes in Phys. 195, Springer-Verlag, Berlin, New York, 1984, pp. 384–412.
[44] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh, 115 (1990), pp. 193–230.
[45] L. Tartar, *On mathematical tools for studying partial differential equations of continuum physics: H-measures and Young measures*, in Developments in Partial Differential Equations and Applications to Mathematical Physics, G. Buttazzo, G. P. Galdi, and L. Zanghirati, eds., Plenum, New York, 1991.
[46] L. Tartar, *Some remarks on separately convex functions*, in Microstructure and Phase Transitions, D. Kinderlehrer, R. D. James, M. Luskin, and J. L. Ericksen, eds., IMA Vol. Math. Appl. 54, Springer-Verlag, New York, 1993, pp. 191–204.
[47] L. C. Young, *Lectures on Calculus of Variations and Optimal Control Theory*, W. B. Saunders, Philadelphia, PA, 1969.

SIAM J. MATH. ANAL.
Vol. 30, No. 6, pp. 1391–1400

ASYMMETRIC MODES IN SYMMETRIC NONLINEAR OPTICAL WAVEGUIDES*

DAVID ARCOYA†, SILVIA CINGOLANI‡, AND JOSÉ L. GÁMEZ†

Abstract. We study a symmetric nonlinear value problem in all $\mathbf{R}$, arising in nonlinear optics from the study of propagation of electromagnetic guided waves through a layered medium with a nonlinear response. By variational arguments, we prove the existence of a positive asymmetric solution of the problem, corresponding to an asymmetric guided wave.

Key words. electromagnetic guided wave, nonlinear eigenvalue problem, asymmetric positive solution

AMS subject classifications. 34B15, 35Q60, 49R05

PII. S0036141098336388

1. Introduction.

The propagation of electromagnetic guided waves through a medium consisting of three layers of dielectric materials has been studied in several papers; see, for instance, [1, 13, 2]. The ability of such slab geometry to support guided waves depends upon the way in which the refractive index varies across the layers. For example, in the absence of nonlinear effects, the condition for guidance requires that the refractive index in the outer layers be smaller than that in the central region, as one might guess from Snell's law. As discussed in [1, 13], nonlinear effects can be used to obtain guidance properties.

In this paper, we consider the case where the medium is stratified in three layers of homogeneous composition perpendicular to the x-axis. In such a medium we seek solutions of Maxwell's equations corresponding to an electric field which is monochromatic, propagating in the direction z and polarized along the y-axis. A field of this kind is given by $E(x,y,z,t) = u(x)\cos(\beta z - \omega t)e_2$, where $\beta > 0$, e_i denotes the usual basis vector and $u : \mathbb{R} \to \mathbb{R}$. This ansatz leads to a second-order nonlinear eigenvalue problem on the real axis:

$$-\ddot{u}(x) + \beta^2 u(x) = \frac{\omega}{c^2} n^2 \left(x, \frac{1}{2}u^2(x) \right) u(x) \quad \text{for } x \in \mathbb{R},$$

where c is the speed of light in the vacuum and $n^2(x,s)$ is the dielectric function.

The guidance conditions require that all fields decay to zero as $|x| \to +\infty$ and that the total electromagnetic energy for unit length in z is finite in each plane $y \equiv$ const. This amounts to

$$\lim_{|x|\to\infty} u(x) = \lim_{|x|\to\infty} \dot{u}(x) = 0$$

*Received by the editors March 31, 1998; accepted for publication (in revised form) December 8, 1998; published electronically October 13, 1999.
http://www.siam.org/journals/sima/30-6/33638.html

†Departamento de Análisis Matemático, Universidad de Granada, 18071-Granada, Spain (darcoya@goliat.ugr.es, jlgamez@goliat.ugr.es). These authors were supported by D.G.E.S. Ministerio de Educación y Ciencia (Spain) and Acción Integrada Spain-Italy HI1997-0049 and by E.E.C. contract ERBCHRXCT940494.

‡Dipartimento di Matematica, Politecnico di Bari, 70125-Bari, Italy (cingolan@pascal.dm.uniba.it). This author was supported by M.U.R.S.T. (40% and 60% funds) and E.E.C. program Human Capital Mobility contract ERBCHRXXCT 9400494.

and

$$\int_{\mathbb{R}} u^2(x)\,dx + \int_{\mathbb{R}} \dot{u}^2(x)\,dx < \infty.$$

This problem has been studied for special choices of the dielectric function n^2. In [1], the dielectric function is taken of the form

$$n^2(x,s) = \begin{cases} q^2 + c^2 & \text{if } |x| < d, \\ q^2 + s & \text{if } |x| > d, \end{cases}$$

where $q, c \in \mathbb{R}$ and $d > 0$ denotes the thickness of the internal layer. In such a case the equation can be integrated directly, and taking $\lambda = \beta^2$ as a bifurcation parameter, one can obtain a family of asymmetric solutions bifurcating from the branch of symmetric ones, at a certain value $\lambda = \lambda_0$, yielding the existence of asymmetric bound states for any $\lambda > \lambda_0$. (See also [5, 6] for discussions about stability.)

In a recent work [2], a class of equations which are not explicitly integrable is considered and a perturbative method is applied to yield existence of asymmetric guided waves when the internal layer of the medium is thin.

In this paper, our goal is to show how variational techniques can be applied to prove existence of asymmetric modes without restrictive assumptions on the thickness of the internal layer.

Precisely, we study a differential equation of the type

$$\begin{gathered} -\ddot{u}(x) + (\lambda - c^2 h(x))u(x) = b(x)|u(x)|^{p-2}u(x) \quad \text{for } x \in \mathbb{R}, \\ u \in H^1(\mathbb{R}), \end{gathered} \tag{1.1}$$

where $\lambda > 0$, $p > 2$, and h, b are even functions such that the supports of h and $1 - b$ are compact.

We point out that (1.1) fits into the Akhmediev setting provided that $\lambda = \beta^2$, $p = 4$, $h = \chi$, $b = 1 - \chi$, with $\chi(x)$ being the characteristic function of $(-d, d)$.

Specifically, we seek positive solutions of (1.1), under the following restrictions on $h, b : \mathbb{R} \to \mathbb{R}$:

(H1) h is a bounded function with $h \geq 0$, $h \not\equiv 0$, supp $h \subset [-1, 1]$;

(B1) b is a bounded function with $b(x) = 1$ for $x \notin [-1, 1]$, $b(x) \leq 0$ for $x \in (-1, 1)$.

We remark that, by making a simple change of variables, the role of the interval $[-1, 1]$ in the assumptions (H1) and (B1) can be played by any closed and bounded interval.

Our variational approach consists of minimizing the Euler functional in a suitable C^1-manifold. In order to do this, we need to check the compactness condition of Palais–Smale, which is obtained by following closely the arguments in [8] and [3]. (See also [9, 10].) By testing the levels of the possible symmetric solutions, we infer that for λ sufficiently large, the minimizer is asymmetric. Similar arguments are used in [4] in a different setting.

Precisely, our result is considered in the following theorem.

THEOREM 1.1. *Assume* (H1) *and* (B1). *If* $\lambda > c^2\|h\|_\infty$, *then the problem* (1.1) *has, at least, one asymmetric positive solution.*

The paper is organized as follows: section 2 is devoted to the proof of the Palais–Smale condition. The details of the minimization and the proof that minimizers are asymmetric are given in section 3.

2. The Palais–Smale condition. In order to prove these technical results, we need only assume the following (less restrictive) conditions on h and b:

(H2) h is a bounded function with $h \geq 0$, $\lim_{|x|\to\infty} h(x) = 0$;

(B2) b is a bounded function with $\limsup_{|x|\to\infty} b(x) = 1$.

Since h is a bounded function, the first eigenvalue $\lambda_1(-c^2 h)$ of the weighted eigenvalue problem associated with the Laplacian operator of weight $-c^2 h$ is given by

$$-c^2\|h\|_\infty \leq \lambda_1(-c^2 h) = \inf_{u\neq 0} \frac{\int_{\mathbb{R}} |\dot{u}|^2\,dx - c^2\int_{\mathbb{R}} h(x)|u|^2\,dx}{\int_{\mathbb{R}} u^2\,dx}.$$

Moreover, assume

$$\lambda > -\lambda_1(-c^2 h) \geq 0, \tag{2.1}$$

and set $q(x) := \lambda - c^2 h(x)$; then it is easy to see that

$$\|u\|^2 := \int_{\mathbb{R}} |\dot{u}|^2 + q(x)|u|^2\,dx, \qquad u \in H^1(\mathbb{R})$$

defines a norm on $H^1(\mathbb{R})$ which is equivalent to the usual norm $\|u\|_{H^1}^2 = \int_{\mathbb{R}} \left(|\dot{u}|^2 + |u|^2\right) dx$. Indeed, we can observe that

$$\|u\|^2 \leq \max\{\|q\|_\infty, 1\}\|u\|_{H^1}^2.$$

Otherwise, for k small enough, we get for any $u \in H^1(\mathbb{R})$,

$$\begin{aligned} \|u\|^2 - k\|u\|_{H^1}^2 &= (1-k)\int_{\mathbb{R}} \left(|\dot{u}|^2 - c^2 h u^2\right)\,dx \\ &\quad + \int_{\mathbb{R}} \left(\lambda - k - kc^2 h\right) u^2\,dx \\ &\geq \left[(1-k)\lambda_1(-c^2 h) + \lambda - k - kc^2\|h\|_\infty\right]\int_{\mathbb{R}} u^2\,dx \\ &\geq 0. \end{aligned} \tag{2.2}$$

From (2.2), it is easy to deduce that $\|u\| = 0$ implies $u = 0$. In addition, $\|\cdot\|$ satisfies the other properties of a norm and thus it is a norm equivalent to $\|\cdot\|_{H^1}$.

Let us consider $J : H^1(\mathbb{R}) \to \mathbb{R}$ the Euler functional associated with the problem (1.1), namely,

$$J(u) = \frac{1}{2}\int_{\mathbb{R}} |\dot{u}|^2 + q(x)|u|^2\,dx - \frac{1}{p}\int_{\mathbb{R}} b(x)|u|^p\,dx,\ u \in H^1(\mathbb{R}).$$

We now prove that the Euler functional satisfies the Palais–Smale condition on a certain sublevel. Let us define

$$m(\lambda) := \inf_{u\in H^1(\mathbb{R})\setminus\{0\}} \frac{\int_{\mathbb{R}} |\dot{u}|^2 + \lambda|u|^2}{\left(\int_{\mathbb{R}} |u|^p\right)^{2/p}}. \tag{2.3}$$

LEMMA 2.1. *Assume* (H2), (B2) *and* (2.1). *Then J satisfies the Palais–Smale condition on the sublevel $\Sigma := \{u \in H^1(\mathbb{R}) : J(u) < \frac{p-2}{2p}\, m(\lambda)^{p/(p-2)}\}$; that is, if u_n is a sequence in $H^1(\mathbb{R})$, such that*

(i) $J(u_n) \to c$,
(ii) $J'(u_n) \to 0$,
with $c < L \equiv \frac{p-2}{2p}\, m(\lambda)^{p/(p-2)}$, *then* u_n *admits a convergent subsequence.*

Proof. We follow closely the arguments in [8] and [3]. Let $\{u_n\}$ in $H^1(\mathbb{R})$ such that

$$J(u_n) = c + o(1), \qquad J'(u_n) = o(1) \quad \text{in} \quad H^{-1}, \tag{2.4}$$

as $n \to \infty$, and assume $c < L$. From (2.4), it follows that

$$\frac{p-2}{2p}\, \|u_n\|^2 \le c + |\langle o(1), u_n\rangle|,$$

and thus $\{u_n\}$ is bounded in $H^1(\mathbb{R})$ and, up to a subsequence, $\{u_n\}$ has a weak limit $u \in H^1(\mathbb{R})$. In order to show that $\{u_n\}$ converges to u strongly in $H^1(\mathbb{R})$, it suffices to prove that

$$\|u_n\| \to \|u\| \qquad \text{as} \quad n \to \infty.$$

Now we observe that for any $R > 0$, there results

$$\begin{aligned} \left| \|u_n\|^2 - \|u\|^2 \right| \le & \left| \int_{|x|\le R} \left(|\dot u_n|^2 + q(x)|u_n|^2 \right) - \int_{|x|\le R} \left(|\dot u|^2 + q(x)|u|^2 \right) \right| \\ & + \int_{|x|>R} \left(|\dot u_n|^2 + |q(x)||u_n|^2 \right) + \int_{|x|>R} \left(|\dot u|^2 + |q(x)||u|^2 \right). \end{aligned}$$

Therefore since the Sobolev imbedding is compact on bounded sets, it suffices to show that for any $\delta > 0$ there exists $R > 0$ such that for any $n \ge R$ there results

$$\int_{|x|\ge R} \left(|\dot u_n|^2 + |q(x)||u_n|^2 \right) < \delta$$

and then, by the weak lower semicontinuity of the above integral,

$$\int_{|x|\ge R} \left(|\dot u|^2 + |q(x)||u|^2 \right) < \delta.$$

By contradiction, assume that there exists δ_0 such that for any $R > 0$ there results

$$\int_{|x|\ge R} \left(|\dot u_n|^2 + |q(x)||u_n|^2 \right) \ge \delta_0$$

for some $n = n(R) \ge R$. As a consequence, there exists a subsequence $\{u_{n_k}\}$ such that

$$\int_{|x|\ge k} \left(|\dot u_{n_k}|^2 + |q(x)||u_{n_k}|^2 \right) \ge \delta_0 \tag{2.5}$$

for any $k \in \mathbf{N}$. For any $r > 0$, let us introduce the annulus

$$A_r = \left\{ x \in \mathbb{R} \ : \ r \le |x| \le r + 1 \right\}.$$

CLAIM. *For any $\xi > 0$ and for any $R > 0$ there exists $r > R$ such that*

$$\int_{A_r} \Big(|\nabla u_{n_k}|^2 + |q(x)||u_{n_k}|^2 \Big) < \xi \tag{2.6}$$

for infinitely many $k \in \mathbf{N}$.

By contradiction, assume that for some $\xi_0, R_0 > 0$ and for any integer $m \geq [R_0]$ there exists $\nu(m) \in \mathbf{N}$ such that

$$\int_{A_m} \Big(|\dot{u}_{n_k}|^2 + |q(x)||u_{n_k}|^2 \Big) \geq \xi_0$$

for any $k \geq \nu(m)$. Plainly, we can assume that the sequence $\{\nu(m)\}$ is nondecreasing. Therefore, for any integer $\overline{m} \geq [R_0]$, there exists an integer $\nu(\overline{m})$ such that

$$\begin{aligned} \int_{\mathbb{R}} \Big(|\dot{u}_{n_k}|^2 + |q(x)||u_{n_k}|^2 \Big) &\geq \int_{[R_0] \leq |x| \leq \overline{m}} \Big(|\dot{u}_{n_k}|^2 + |q(x)||u_{n_k}|^2 \Big) \\ &\geq (\overline{m} - [R_0])\xi_0 \end{aligned}$$

for any $k \geq \nu(\overline{m})$. This contradicts

$$\begin{aligned} \int_{\mathbb{R}} \Big(|\dot{u}_{n_k}|^2 + |q(x)||u_{n_k}|^2 \Big) &\leq \max\{1, \|q\|_\infty\} \int_{\mathbb{R}} \Big(|\dot{u}_{n_k}|^2 + |u_{n_k}|^2 \Big) \\ &\leq K \int_{\mathbb{R}} \Big(|\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2 \Big) \leq K_1 \end{aligned}$$

with K, K_1 positive constants, and it proves the claim.

Now, let $\xi > 0$ be fixed such that $\lambda - \xi > \frac{\lambda}{2} > 0$. Taking into account (H2) and (B2), there exists $R(\xi) > 0$ such that

$$q(x) \geq \lambda - \xi \quad \text{for any } |x| \geq R(\xi), \tag{2.7}$$

$$b(x) \leq 1 + \xi \quad \text{for any } |x| \geq R(\xi). \tag{2.8}$$

Let $r = r(\xi) > R(\xi)$ be as in (2.6), and let $A = A_r$; up to a subsequence, there results

$$\int_A \Big(|\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2 \Big) < \xi \tag{2.9}$$

for any $k \in \mathbf{N}$. Now let us choose any function $\rho \in C^\infty(\mathbb{R}, [0,1])$ such that $\rho(x) = 1$ for $|x| \leq r$, $\rho(x) = 0$ for $|x| \geq r + 1$, and $|\dot{\rho}(x)| \leq 2$ for any $x \in \mathbb{R}$. For any $k \in \mathbf{N}$, let $v_k = \rho u_{n_k}$ and $w_k = (1 - \rho)u_{n_k}$. It is not difficult to see that

$$|\langle J'(u_{n_k}), v_k \rangle - \langle J'(v_k), v_k \rangle| \leq C_1 \xi, \tag{2.10}$$

$$|\langle J'(u_{n_k}), w_k \rangle - \langle J'(w_k), w_k \rangle| \leq C_2 \xi, \tag{2.11}$$

where C_1 and C_2 are positive constants which do not depend on r. First, we prove (2.10):

$$\Big| \langle J'(u_{n_k}), v_k \rangle - \langle J'(v_k), v_k \rangle \Big|$$

$$
\begin{aligned}
&= \Big| \int_{\mathbb{R}} \big(\dot{w}_k \dot{v}_k + q(x) w_k v_k\big) + \int_{\mathbb{R}} b(x) \big(|v_k|^p - |u_{n_k}|^{p-1} v_k\big) \Big| \\
&= \Big| \int_A \rho(1-\rho)\big(|\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2\big) + \int_A \dot{\rho}(1-2\rho) u_{n_k} \dot{u}_{n_k} \\
&\quad - \int_A \dot{\rho}^2 |u_{n_k}|^2 + \int_A b(x)\rho(\rho^{p-1}-1)|u_{n_k}|^p \Big| \\
&\le 2\int_A \big(|\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2\big) + 6\int_A |\dot{u}_{n_k} u_{n_k}| + 4\int_A |u_{n_k}|^2 \\
&\quad + 2\|b\|_\infty \|u_{n_k}\|_\infty^{p-2} \int_A |u_{n_k}|^2 \le 2\int_A \big(|\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2\big) \\
&\quad + 6\left(\int_A |\dot{u}_{n_k}|^2\right)^{1/2} \left(\int_A |u_{n_k}|^2\right)^{1/2} + M\int_A |u_{n_k}|^2 \\
&\le 2\int_A \big(|\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2\big) \\
&\quad + 6\left(\int_A |\dot{u}_{n_k}|^2 + q(x)|u_{n_k}|^2\right)^{1/2} \left(\int_A |u_{n_k}|^2\right)^{1/2} + M\int_A |u_{n_k}|^2.
\end{aligned}
$$

By (2.7) and (2.9), we infer

$$(\lambda - \xi)\int_A |u_{n_k}|^2 \le \int_A q(x)|u_{n_k}|^2 < \xi,$$

and thus

$$\int_A |u_{n_k}|^2 < \frac{\xi}{(\lambda-\xi)} < \frac{2\xi}{\lambda}.$$

Hence we deduce that

$$\Big|\langle J'(u_{n_k}), v_k\rangle - \langle J'(v_k), v_k\rangle\Big| < 2\xi + 6\xi^{1/2}\left(\frac{\xi}{\lambda}\right)^{1/2} + 2M\frac{\xi}{\lambda} \le C_1\xi.$$

Similar arguments show that (2.11) holds. By (2.10) and (2.11), we deduce

$$o(1) = \langle J'(v_k), v_k\rangle + O(\xi), \qquad o(1) = \langle J'(w_k), w_k\rangle + O(\xi),$$

whence

$$\|v_k\|^2 = \int_{\mathbb{R}} b(x)|v_k|^p + O(\xi), \tag{2.12}$$

$$\|w_k\|^2 = \int_{\mathbb{R}} b(x)|w_k|^p + O(\xi). \tag{2.13}$$

By (2.7) and (2.13), we deduce

$$
\begin{aligned}
c + o(1) &= J(u_{n_k}) = J(v_k) + J(w_k) + O(\xi) \\
&\ge \frac{p-2}{2p}\int_{\mathbb{R}} \Big(|\dot{w}_k|^2 + \lambda|w_k|^2\Big) + O(\xi).
\end{aligned} \tag{2.14}
$$

By (2.5) and (2.8), we infer

$$\begin{aligned}\int_{\mathbb{R}} |w_k|^p &\geq \int_{\mathbb{R}} b(x)|w_k|^p + O(\xi) = \|w_k\|^2 + O(\xi)\\ &\geq \int_{|x|\geq r+1} \Big(|\dot{w}_k|^2 + q(x)|w_k|^2\Big) + O(\xi) \geq \frac{1}{2}\delta_0\end{aligned}$$

for ξ small and k large. Hence for ξ small,

$$\begin{aligned}\int_{\mathbb{R}} \Big(|\dot{w}_k|^2 + \lambda|w_k|^2\Big) &\geq m(\lambda)\Big(\int_{\mathbb{R}} |w_k|^p\Big)^{2/p}\\ &\geq m(\lambda)\Big(\int_{\mathbb{R}} \Big(|\dot{w}_k|^2 + \lambda|w_k|^2\Big) + O(\xi)\Big)^{2/p},\end{aligned}$$

and thus

$$\int_{\mathbb{R}} \Big(|\dot{w}_k|^2 + \lambda|w_k|^2\Big) \geq m(\lambda)^{p/(p-2)} + \frac{O(\xi)}{\Big(\int_{\mathbb{R}} |\dot{w}_k|^2 + \lambda|w_k|^2\Big)^{2/(p-2)}}.$$

Hence from (2.14), we infer

$$c + o(1) \geq \frac{p-2}{2p}\, m(\lambda)^{p/(p-2)} + O(\xi)$$

for ξ small and k large. Letting $k \to \infty$ and $\xi \to 0$ yields a contradiction and concludes the proof. □

3. Existence of asymmetric solution. Let us consider the C^1-manifold

$$M = \left\{ u \in H^1(\mathbb{R}) \,:\, \int_{\mathbb{R}} b(x)|u|^p\, dx = 1 \right\}.$$

As $J(u) = \frac{1}{2}\|u\|^2 - \frac{1}{p}$, for any $u \in M$, the minimization of $J_{|_M}$ is equivalent to the minimization of the functional $I : M \to \mathbb{R}$ defined by

$$I(u) = \|u\|^2, \quad u \in M.$$

LEMMA 3.1. *Assume* (2.1). *Then we have the following.*

(i) *M is a closed C^1-manifold of codimension one.*

(ii) *For any $v \in H^1(\mathbb{R})$ satisfying $\int_{\mathbb{R}} b(x)|v|^p\, dx \neq 0$, there exists a unique $\alpha > 0$ such that $\alpha v \in M$. Indeed,*

$$\alpha = \left(\int_{\mathbb{R}} b(x)|v|^p\, dx\right)^{-1/p} \quad \text{and} \quad I(\alpha v) = \frac{\|v\|^2}{\big(\int_{\mathbb{R}} b(x)|v|^p\, dx\big)^{2/p}}.$$

Proof. (i) Let $f(u) = \int_{\mathbb{R}} b(x)|u|^p\, dx$, $u \in H^1(\mathbb{R})$. Observe that for any $u \in M$, $f(u) = 1$; hence

$$\langle f'(u), u\rangle = -p \int_{\mathbb{R}} b(x)|u|^p\, dx = -p. \tag{3.1}$$

As a consequence, $f'(u) \neq 0$ if $u \in M$. This means that M is a C^1-manifold of codimension one. Trivially, it is closed by the continuity of the map $u \mapsto \int_{\mathbb{R}} b(x)|u|^p\, dx$.

(ii) Straightforward computations show this part. □

LEMMA 3.2. *Assume* (H1), (B1), *and* (2.1). *Then*

$$\inf_{v\in M} I(v) < m(\lambda).$$

Proof. By [12, Theorem 5.5], it is known that the infimum in (2.3) is attained at a symmetric function $z \in H^1(\mathbb{R})$ which is a solution of the problem

$$\begin{gathered} -\ddot{u}(x) + \lambda u(x) = |u(x)|^{p-2}u(x) \quad \text{for } x \in \mathbb{R}, \\ \lim_{|x|\to\infty} u(x) = 0. \end{gathered}$$

In addition, by [7, 11], such a function is unique and it is given by

$$z(x) = \left[\frac{p\lambda}{2\cosh^2(\frac{p-2}{2}\sqrt{\lambda}x)}\right]^{1/(p-2)}, \quad x \in \mathbb{R}.$$

For $\theta \in \mathbb{R}$, take the translate $z_\theta = z(\cdot + \theta)$ and consider, by (ii) of the preceding lemma, the positive number α_θ such that $\alpha_\theta z_\theta \in M$. The proof would be concluded if we show that $I(\alpha_\theta z_\theta) < m(\lambda)$ for some θ. In order to verify this, observe that from hypotheses (H1) and (B1), the existence of the positive limit of the function $z(x)e^{\sqrt{\lambda}x}$ as $x \to +\infty$ implies that for large θ,

$$\int_{\mathbb{R}} h(x)z_\theta^2 dx \geq C_1 e^{-2\sqrt{\lambda}\theta} \quad \text{and} \quad \int_{\mathbb{R}} (1-b(x))|z_\theta|^p dx \leq C_2 e^{-p\sqrt{\lambda}\theta}$$

for some positive constants C_1 and C_2. Then,

$$\begin{aligned} I(\alpha_\theta z_\theta) &= \frac{\displaystyle\int_{\mathbb{R}} (|\dot{z}|^2 + \lambda z^2)\,dx - c^2 \int_{\mathbb{R}} h(x) z_\theta^2\,dx}{\displaystyle\left[\int_{\mathbb{R}} |z|^p\,dx - \int_{\mathbb{R}} (1-b(x))|z_\theta|^p\,dx\right]^{2/p}} \\ &\leq \frac{\displaystyle\int_{\mathbb{R}} (|\dot{z}|^2 + \lambda z^2)\,dx - c^2 C_1 e^{-2\sqrt{\lambda}\theta}}{\displaystyle\left[\int_{\mathbb{R}} |z|^p\,dx - C_2 e^{-p\sqrt{\lambda}\theta}\right]^{2/p}} \\ &= \frac{\displaystyle\int_{\mathbb{R}} (|\dot{z}|^2 + \lambda z^2)\,dx}{\displaystyle\left[\int_{\mathbb{R}} |z|^p\,dx\right]^{2/p}} \; \frac{1 - C_3 e^{-2\sqrt{\lambda}\theta}}{\left(1 - C_4 e^{-p\sqrt{\lambda}\theta}\right)^{2/p}} \\ &= m(\lambda)\; \frac{1 - C_3 e^{-2\sqrt{\lambda}\theta}}{\left(1 - C_4 e^{-p\sqrt{\lambda}\theta}\right)^{2/p}}. \end{aligned}$$

Since $p > 2$, for large values of θ, one has $(1 - C_3 e^{-2\sqrt{\lambda}\theta})/(1 - C_4 e^{-p\sqrt{\lambda}\theta})^{2/p} < 1$ and consequently, $I(\alpha_\theta z_\theta) < m(\lambda)$. □

Now we are ready to prove Theorem 1.1.

Proof of Theorem 1.1. The solution of (1.1) is found as a minimizer of I constrained on M (up to a Lagrangian multiplier). In order to prove the existence of such a minimizer, and taking into account Lemmas 3.1 and 3.2, it suffices to show that I satisfies the Palais–Smale condition on the sublevel $M_\lambda := \{u \in M : I(u) < m(\lambda)\}$, that is, if $\{u_n\}$ is a sequence in M, such that

(a) $I(u_n) \to c$,
(b) $|I'(u_n)(v)| \le \varepsilon_n \|v\|$ for any $v \in T_{u_n}M$, with $\varepsilon_n \to 0$,
with $c < m(\lambda)$, then $\{u_n\}$ admits a convergent subsequence. Indeed, by (b), it follows that there exists a sequence $\{\theta_n\} \subset \mathbb{R}$ such that for any $n \in \mathbf{N}$

$$-\triangle u_n + q(x)u_n - \theta_n b(x)|u_n|^{p-2}u_n = o(1) \quad \text{in} \quad H^{-1}.$$

Testing the previous equation in u_n gives

$$\theta_n = I(u_n) - \langle o(1), u_n \rangle.$$

Therefore, setting $v_n(x) = I(u_n)^{1/(p-2)}\ u_n$, we plainly get

$$J'(v_n) = I(u_n)^{1/(p-2)} o(1) = o(1) \quad \text{in} \quad H^{-1}.$$

Moreover there results

$$J(v_n) = \frac{p-2}{2p} I(u_n)^{p/(p-2)} = c' + o(1)$$

with $c' < \frac{p-2}{2p} m(\lambda)^{p/(p-2)}$. Therefore, by Lemma 2.1, it follows that $\{v_n\}$ is precompact, and thus, taking into account (a), $\{u_n\}$ is precompact and the Palais–Smale condition holds.

Therefore, if v is a minimizer of $I_{|_M}$, setting $\overline{u} = I(v)^{1/(p-2)}v$, it results that $\overline{u}$ is a solution of (1.1). Moreover $J(\overline{u}) = J\big(I(v)^{1/(p-2)}v\big) < \frac{p-2}{2p} m(\lambda)^{p/(p-2)}$.

Therefore, the proof would be concluded if we show that $\overline{u}$ is not symmetric for $\lambda > c^2\|h\|_\infty$. This is a consequence of the fact that every possible symmetric positive solution u of (1.1) must have a level greater than $\frac{p-2}{2p}\, m(\lambda)^{p/(p-2)}$. Indeed, since $b(x) \le 0$ for any $x \in (-1,1)$, for such u,

$$-\ddot{u}(x) + (\lambda - c^2 h(x))u(x) \le 0 \quad \text{for } x \in (-1,1).$$

Hence, since $\lambda > c^2\|h\|_\infty$, we get $\ddot{u}(x) \ge (\lambda - c^2h(x))u(x) \ge 0$ for any $x \in (-1,1)$. Then u is convex in $(-1,1)$, and thus $\max\{u(x) : x \in [-1,1]\} = u(-1) = u(1)$, with $\dot{u}(1) > 0$, by the Hopf lemma. This means that the maximum of u in all $\mathbb{R}$ is attained in some point $x_0 \in \mathbb{R} - [-1,1]$.

Now, observing that supp $h \subset [-1,1]$, we obtain $-\ddot{u}(x) + \lambda u(x) = u(x)^{p-1}$, for $|x| > 1$, which together with the fact $\dot{u}(\pm x_0) = 0$ implies that

$$u(x) = z(|x| - x_0) \quad \text{for any } |x| \ge x_0.$$

As a consequence, since $\lambda > c^2\|h\|_\infty$, and $-\ddot{z} + \lambda z = z^{p-1}$, we obtain

$$\begin{aligned}
\frac{2p}{p-2}\ J(u) &= \int_{\mathbb{R}} \left[|\dot{u}|^2 + (\lambda - c^2h)u^2\right] dx \\
&\ge \int_{\mathbb{R}\setminus[-x_0,x_0]} \left[|\dot{u}|^2 + (\lambda - c^2h)u^2\right] dx \\
&= \int_{\mathbb{R}\setminus[-x_0,x_0]} |\dot{u}|^2 + \lambda u^2\, dx \\
&= \int_{\mathbb{R}} |\dot{z}|^2 + \lambda z^2\, dx
\end{aligned}$$

$$
\begin{aligned}
&= \left[\frac{\int_{\mathbb{R}} |\dot{z}|^2 + \lambda z^2 \, dx}{\left(\int_{\mathbb{R}} |\dot{z}|^2 + \lambda z^2 \, dx \right)^{2/p}} \right]^{p/(p-2)} \\
&\geq \left[\frac{\int_{\mathbb{R}} |\dot{z}|^2 + q(x)|z|^2 \, dx}{\left(\int_{\mathbb{R}} |z|^p \, dx \right)^{2/p}} \right]^{p/(p-2)} \\
&= m(\lambda)^{p/(p-2)},
\end{aligned}
$$

which concludes the proof. □

REFERENCES

[1] N. N. Akhmediev, *Novel class of nonlinear surface waves: Asymmetric modes in a symmetric layered structure*, Sov. Phys. JEPT, 56 (1982), pp. 299–303.

[2] A. Ambrosetti, D. Arcoya, and J. L. Gámez, *Asymmetric bound states of differential equations in nonlinear optics*, Rend. Sem. Mat. Univ. Padova, 100 (1998), pp. 231–247.

[3] S. Cingolani and M. Lazzo, *Multiple semiclassical standing waves for a class of nonlinear Schrödinger equations*, Topol. Methods Nonlinear Anal., 10 (1997), pp. 1–13

[4] M. J. Esteban, *Nonsymmetric ground states of symmetric variational problems*, Comm. Pure Appl. Math., 44 (1991), pp. 259–274.

[5] M. Grillakis, J. Shatah, and W. Strauss, *Stability theory of solitary waves in the presence of symmetry*, J. Funct. Anal., 74 (1987), pp. 160–197.

[6] C. K. R. T. Jones and J. V. Moloney, *Instability of standing waves in nonlinear optical waveguides*, Phys. Lett. A, 117 (1986), pp. 176–184.

[7] M. K. Kwong, *Uniqueness of positive solutions of* $\Delta u - u + u^p = 0$ *in* R^n, Arch. Rational Mech. Anal., 105 (1989), pp. 243–266.

[8] Y. Y. Li, *Existence of multiple solutions of semilinear equations in* $\mathbb{R}^n$, Progr. Nonlinear Differential Anal., 4 (1990), pp. 134–159.

[9] P. L. Lions, *The concentration-compactness method in the calculus of variations: The locally compact case. Part* I, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), 109–145.

[10] P. L. Lions, *The concentration-compactness method in the calculus of variations. Part* II, Ann. Inst. H. Poincaré Anal. Non Linéaire, 1 (1984), pp. 223–283.

[11] K. Mcleod and J. Serrin, *Uniqueness of positive radial solution of* $\Delta u + f(u) = 0$ *in* $\mathbb{R}^n$, Arch. Rational Mech. Anal., 99 (1987), pp. 115–145.

[12] C. Stuart, *Bifurcation in* $L^p(\mathbb{R}^N)$ *for a semilinear elliptic equation*, Proc. London Math. Soc. (3), 57 (1988), pp. 511–541.

[13] C. Stuart, *Guidance properties of nonlinear planar waveguided*, Arch. Rational Mech. Anal., 125 (1993), pp. 145–200.